EUROPEAN UNIVERSITY INSTITUTE
Department of Political and Social Sciences

- The One, The Two, and The Many - Autonomous, self-interested, and interacting individuals in images of social order within contemporary social thought

by

Celia de Andrade Lessa

Thesis submitted for assessment with
a view to obtaining the Degree of Doctor of the
European University Institute

Florence, March 1998

EUROPEAN UNIVERSITY INSTITUTE

Department of Political and Social Sciences

- The One, The Two, and The Many - Autonomous, self-interested, and interacting individuals in images of social order within contemporary social thought

by

Celia de Andrade Lessa

Thesis submitted for assessment with
a view to obtaining the Degree of Doctor of the
European University Institute

Examining jury:
Prof. Colin Crouch (EUI)
Prof. Jean-Pierre Dupuy (CREA - Paris)
Prof. Alan Kirman (GREQAM - Marseille)
Prof. Steven Lukes (Università di Siena - Supervisor)

Florence, March 1998

To Jaques

and Antonio,

in praise of love and hope.

CONTENTS

PART II

AUTONOMY - BEYOND SELF-INTEREST

(THE ONE)

PART III

INTERACTING INDIVIDUALS

(THE MANY)

Acknowledgments

A number of people helped me in many different ways with my PhD research. I should like to mention in the first place my supervisor, Professor Steven Lukes, who was always available to meet me on every important occasion, and whose advice I benefited from a lot. He was also very supportive (and patient) in backing my leavings and coming backs, for on two occasions I had to interrupt the PhD research to resume my teaching duties in Brazil. More than that, on every such occasion he always displayed a confidence in my work that functioned, I can see it now, in the manner of a self-fulfilling prophecy, which encouraged me further to keep going after more than one adversity. In still another important sense, I am grateful to him as a teacher for I had the privilege of attending his weekly seminars in which much of my own intellectual outlook was shaped in the light of his well-known intellectual qualities.

I am also very grateful to my former masters, Professor Wanderley Guilherme dos Santos and Maria da Conceição Tavares, whose academic freedom and creativity have left indelible marks in my spirit. I am indebted to a great extent to Professor Antonio Barros de Castro for his sensitivity, as he was the first to encourage me to follow the road of ideas.

To Professor César Guimarães, my Brazilian advisor of terrible intelligence and great human qualities, I owe a lot: good chat, good advice, and kindness. Without his personal endeavors I would not have received financial support from our official institutions in a rather crucial moment of my work.

I should like to thank Professor Alan Kirman, for patient advice on game-theoretical and social-choice matters to a non-specialist reader, for reading through my texts and suggesting bibliography and interesting points that I tried to pursue, at least to some extent. I am also indebted to Professor Spyros Vassilakis, whose office I invaded in a moment of hardship, for teaching me a number of game-theoretical matters, and suggesting precious references as well as ideas. Doctor Árpad Szacolczai was very kind in reading the very early sketches of my work and conceding me generous portions of his time. Professor Alessandro Pizzorno also helped me putting the ideas into better shape during a working-lunch. Professor Jean-Pierre Dupuy, in an inspiring meeting over the Arno, listened with attention to my exposition of an early version of

the original project and suggested helpful bibliographical references. Also his suggestions were very helpful at a later stage as they brought to my attention the need for important changes to the general structure of the thesis.

I am also indebted to João Carlos Espada, for letting me have a look at some very interesting pieces of his recent book.

This thesis would not have been possible were it not for the financial support I received, over the period, successively by CAPES, CNPq and the Italian Ministero dell'Estero.

I owe to a number of colleagues at the Department of Economics, in the Universidade Federal Fluminense, the crucial participation in departmental meetings where my leave of absence from the university was at stake. I should mention, in particular, my colleagues and friends Ricardo Henriques, Maria Bernadete Gutierrez, Inês Patricio and Antonio Fiorêncio.

The staffs of the European University Institute and IUPERJ, in Brazil, were always very helpful; without their aid the path would have been much more difficult to follow. Françoise Thauvin, Ken Hulley, Ursula Brose, the secretaries of the SPS department, Eva Breivik, in particular, the very special secretary of the Economics department, Jessica Spataro, the library staff, the porters, the mensa staff, they were all too eager to help on every occasion. I owe many thanks to Nicki Hargraves for engaging in a tandem with me, which, I suspect, was actually motivated by her willingness to help me in developing my skills in English; if any doubt remains, suffice it to say that after a month of language 'exchange' (as most of the time was dedicated to English) she was not able to speak more than two words in Portuguese (which she knew already!) . To Nicky Owtram, for English assistance. Doctor Andreas Frijdal granted my scholarship for two rather crucial moments. On the side of my compatriots, I should mention the library staff of the IUPERJ, in particular, Simone, and Luiza and Lia who helped me in printing.

In Italy, I counted upon the affective support and stimulating intellectual environment provided by many multinational friends. I am not sure whether I will remember them all here as they are so numerous. But I cannot help mentioning my friends Antonio Goucha Soares, Donatella Campus,

Katia Martinez and Simone Frasca (on three occasions my very kind hosts), Per and Anette Mouritsen, Valérie Campos Mello, Anton Schutz and Chantal Miller, Marcelo Oviedo and Debra, Maria Pinto, Amy Verdun and Alvaro Oliveira. Andréa shared with me the common experience of being a Brazilian woman with a little child to look after far away from home. Pina brought home very near, our Italian 'mamma' and a 'nonna' to my son; she synthesized Italian temperament and generosity in every detail.

Marcos Campos made the entire enterprise possible, as my procurator. My family, my grandmother, my parents, my sister and brothers, rendered the unbearable (for an inhabitant of the tropical side of the world) cold winter days warmer, with their touching letters. Maria Alice Rabello accompanied the enterprise from the birth to the end, and did nothing to stop it, she therefore is partially guilty of it as she convinced me to deviate from the (other) temptations on the way. Bia and Lucia proved that fraternity is not a matter of blood.

My mother was inestimable in many respects. From the beginning, she has been an 'activist' for my academic activities; this time, however, she surpassed herself, in letters, long-distance calls, and eventually in coming personally to Florence, leaving behind her professional activity as a psychoanalyst, for two rather long periods where she looked after my son and taught me how to be a mother. I hope to have learnt the lesson.

My son, Antonio, grew up with the thesis. I came to bring him up together with my getting acquainted with Florence and my looking into the ideas I wanted to pursue; everything together at the same time, they came to be mixed in my mind, the novelties infusing strength in one another, they put me in a state of enthusiasm which may have looked bizarre to those who could not outguess the uniqueness of that moment for me. My love for Antonio pushed me further, his presence is all along.

My dear Jaques, companion of deeds and dreams, covered me with so many things just by way of letting his nature flow, kindness, solidarity, love, tenderness, so many riches brought out without effort, that I suspect there still exists innocence in the world, and things could start anew. My Florentine time was then sweet for I was lucky to have him on my side.

Foreword

Something I have learnt from my teachers is that the hallmark of intellectual virtue is the ability to pose interesting questions. A lesson I have come to understand on my own, however, is that this ability is something we develop over time and with practice, if ever.

As a typical student of political philosophy I began rather inaccurately by taking an unlimited interest in the social world in its infinity, and then, in how it manages to hang together and to survive the immense dangers that beset, and after that, what prospects there are for hope. Still, as such a student, I realized that in order to get some clues to satisfy my curiosity, in order to find my way, I should enter into another world, a world of sophisticated representations of that infinity. Somewhere on the way I figured out that the immense world that was the object of my interest and the sophisticated representations that somehow tried to measure it were not entirely set apart from each other.

Next, an apparently interesting question arose: if, as I was ready to believe, the social world is a world of representations, as people act and react according to the way they re-present the situation in which they themselves happen to be, to what extent are the sophisticated representations (our social philosophy and theories) sensitive and responsive to this 'fact'?

This was my background question and its motivation. Shortly after they were established I went on to select my 'cases' and my 'targets'.

I then decided to concentrate on the so-called individualist tradition, aware, however, of the conceptual difficulty in identifying a clear-cut definition of 'individualism' as Steven Lukes had pointed out in his book Individualism. I dealt with this problem by taking the word of the authors as authoritative of their 'individualisms' and trying to make explicit the kind of individualism they claimed to follow. I felt that the individualist assumption of the precedence of individuals in relation to the social dimension as far as a representation of the social world is concerned was a good starting point just because, as far as I could see, it seemed to be more provocative and challenging a premise than the converse presupposition of an already existing

social dimension. Striving to prove their case individualist theses may display what a non-individualist cannot, or it seemed to me, namely, what a representation in terms of individuals is able to achieve and what it is not able to achieve at all, both in terms of explanatory findings and normative horizons for our social world. Of course, the result is a variegated lot of 'individualist' arguments.

So, in its form, the thesis is an arranged collection of interpretative exercises whose initial task is to discern distinct images of the social world amidst contemporary social thought, as these present the social order in terms of the interaction of individuals. Accordingly, it identifies and subsequently comments on three major varieties of idealized representations of the social order, in recent developments of the so-called individualist tradition. The unifying thread is that these varieties of individualism are reconstructed so as to spell out the kind of **idealization** undertaken, as well as the extent to which, on the way to the ideal, the social dimension is **reduced** to the individual one.

Let me introduce the three varieties by explaining the title. (Note, however, that the order of the exposition is slightly different from the order in the title mainly for rhetoric reasons as explained in the Introductory Chapter.)

Thus, the '**one**' world ('autonomy' or 'beyond self-interest'), is designed so as to include pieces of contemporary thinking, such as Rawls's theory of justice and Harsanyi's rule-utilitarianism, that share the assumption that the social order conceived in terms of individuals may achieve some ideal of well-orderliness. To this end, an assumption of the individuals as **also** moral persons is adopted. As individuals are conceived as moral persons, **one** individual is sufficient to reason to an ideal society; the reduction of the social dimension into the individual one is therefore maximum in this approach.

Conversely, the '**two**' world ('self-interest and beyond'), is organized around the distinct vision that the social order may be thought of exclusively in terms of the self-interest of the individuals. The ideal social order is either conceived as coordination of people's plans or also

efficient cooperation among them. The premise regarding the possibility of reduction is that there is some reality in thinking of at least **two** distinct rationalities or thought processes, not easily reducible to one other, and that this not only matters but there is some sense in trying to model this situation as strategic thinking.

The third framework, the '**many**' or 'interacting individuals' world, is built upon the Hayekian alternative picture that the relationship between the individuals and the social order, at a conceptual level, is complex, the individuals being somehow incomplete in their capacities as sources of order unless they are envisioned as interacting- or rule-guided- individuals. This is not to say that the 'social' is prior to the individual dimension, let alone that it can be accounted for wholly in terms of individuals' conscious actions, as intersubjective rules that are not always consciously followed can account for the complex result which is a society. A more complex view is proposed where the social order surfaces from the interaction of the irreducibly **many**. The ideal here is that some (progressive) order come into existence.

As a final target-related word, my interpretation of this set of theories stresses the fact that reductionism and idealization are in each approach related to its distinctive **knowledge** assumptions. In connection with my initial question as to the theories' sensitivity to our social world being a world of representations, I became convinced that the 'many' outlook is the more sensitive one in that it takes up a less naturalistic stand regarding our knowledge possibilities both as social actors and theoreticians.

Neither the collection of cases nor the way it is arranged, it goes without saying, is intended to be exhaustive. Besides, though arranged in accordance with some common concerns which may suffice to characterize and, to a certain extent, contrast the three accounts here, I have tried to remain sensitive to the internal characteristics of each one, in terms of some of the debates within them and the way these related to my unifying question of how the arguments by the authors examined configure an individualist image of social order, as well as the degree of idealization and reductionism involved. Of course, much of my personal view is revealed by the way I have undertaken to present the distinct perspectives, as sometimes critique appears

much more in its connotation as criticism than as analysis, on one occasion it even amounts to the reconstruction *ab ovo* of a vision from a certain tradition (Part III). In particular, the treatment in Part II is quite conventional as far as Rawls and Harsanyi's theories are concerned, in Part I it is quite unconventional, whereas in Part III I try to rescue the social philosophy of Hayek and reconcile it with non-conservative premises, so that my own 'God's-finger' is much more perceivable.

In between the extreme demands of a thorough collection and those of an idiosyncratic choice, I have worked out an effort of systematization in the Introductory Chapter, where a taxonomy is tried out and the sequence of the chapters is justified.

INTRODUCTORY PART

Introductory Chapter

PICTURES

One may think of social order with the assistance of a number of images, as when one portrays it as somehow resulting from the interaction of its individual members,[1] or else as somehow being the prior background condition for individuals to exist and flourish.[2]

In the individualist literature, the social order is often seen as the outcome of a peculiar **choice** made by rational individuals. In this version, I submit, two kinds of purposes may be supposed of the individuals, as well as two types of rationality. Thus, the ends of the individuals (dimension **E** below) may be regarded, roughly, as **public** or **non-public**, in the sense that the social order may be depicted as the result of individuals' efforts either to achieve it or something else (e.g., their private aims). Similarly, rationality (dimension **R** below) may be seen as **parametric** or **strategic** (to follow Elster's (1979) suggestion), according to whether or not, on the way to furthering their ends, individuals supposedly frame their choice decision in terms of some suitable parameters, external to their deliberation.

We can characterize individuals' coordination with a simple 2x2 matrix, and it takes little time to figure out a number of theories that might illustrate its four cells,[3] providing four different pictures. Provisionally, the following examples chosen from economic, social, and political theories, are suggested:

---

[1] Kukathas and Pettit (1990) suggest the distinction between two varieties of individualism, namely, a 'metaphysical' and a 'moral' individualism. A deep and forceful conceptual discussion of 'individualism' is found in Lukes' (1973) seminal work. Since this is not my main topic of interest here I let the authors speak for themselves about their self-proclaimed 'individualism'.

[2] On some varieties of this image, see Turner (1994).

[3] I disregard, at this point, whether these theories are explanatory or normative.

MATRIX 1

R\E	non-public	public
parametric	**(I)** standard economic theory and conventional welfare economics	**(II)** social choice theory; Rawls & Harsanyi's normative theories
strategic	**(III)** standard game theory	**(IV)** public goods & collective action literature

To a certain extent, each cell displays a model of the coordination of individuals' actions, combining rationality and purpose as variously defined. Economic coordination, to take one example, may conceivably result either from parametric or strategic choices over **non-public** outcomes (where the individuals are not directly interested in their overall coordination), cells I and III respectively, according to whether it is portrayed by the model of a representative individual (who takes the whole as a parameter) or by that of game theory[4] (where the individual relates to some other individual(s) at a local level). In the same way, more general images of **social** coordination might be evoked as illustrations of cells I and III, such as those depicted in the tradition of eighteenth century Scottish thinkers,[5] or else, that which has followed the von Neumann-Morgenstern's game theoretical vision of social interdependence.[6] On the **public** side of the matrix, social coordination may be said to emerge from parametric choices of rational individuals aiming at the very constitution of the social order, as in normative theories where a well-ordered society is the result of a morally constrained individual's rational choice among social states. It may also be said to result from the strategic reasoning of maximizing agents willing to produce it for their own sake, as in versions of

[4] See Kirman (1992) for a discussion of the use of the assumption of the representative individual in economics. See also Kirman (1995) for an alternative evolutionary and game-theoretic approach of economic coordination.

[5] See Hirschman, 1977.

[6] See Leonard, 1995. See also chapter 3.

public-goods literature.

Prima facie, a natural contrast emerges here with respect to the usual explanatory one-dimensional classification of views of social order in terms of causal or intentional explanations, where individuals are thought of either as simple agents to that order or also as effective designers of it. In the latter case, the agent strives for the order and it arises as a result, whereas in the former the agent does not intend the order in his actions though it emerges all the same. This classification seems to coincide with the non-public/public distinction in the matrix above.

On the way to delving into the traditional taxonomy, one might make the rather trivial point that an action always follows an intention (or may be understood in its terms) or else that intentions have an important causal aspect to the extent that they determine the result. In order to avoid a conceptual confusion at this point, it would be useful to somehow restate the distinction between **causal** and **intentional** explanations. To this end, I suggest that in a strictly causal explanation, we the theoreticians, are not supposed to be able to trace all the way from the effects back to the originating causes, as we are not able to predict the course of events from a knowledge of their causes, whereas in an intentional explanation we are supposed to be able to do so in either ways.

Having restated the causal-intentional distinction in the above way, we are now in a position to entertain the rather non nonsensical idea that a cause, in a purely causal explanation, may accommodate an intention on the part of the individual agent, only that this intention is to be systematically **falsified**[7] after all the causes have realized their effects. Conversely, in a strictly intentional explanation the intention is to be **corroborated** when effects are computed. This perception has the merit of highlighting the role of social actors' conjectures and expectations in the social intercourse, however conceived, as crucial intermediaries between intentions and actions. The assumption is that this epistemic material of our actions has an impact over both

[7] Falsified here does not mean 'contradicted' but indicates that there are other things to coordination than individuals' intentions alone, that intentions are 'surpassed' or 'transcended' somehow.

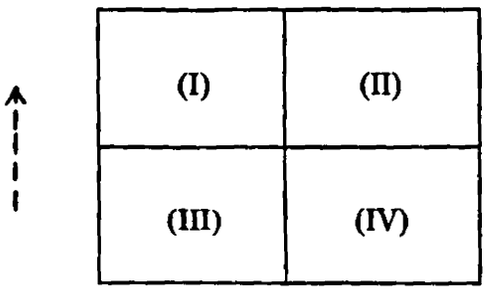
explanatory and normative reach of our theories.

The idea of qualifying causal and intentional models in terms of 'falsification' or 'corroboration', in epistemic terms, that is, acquires more significance in the light of models where social actors are variously thought of in their capacities as both 'consumers' and 'producers' of **knowledge** during social intercourse. This aspect may, to a certain extent at least, be captured through the parametric-strategic distinction. In a typically parametric approach, the decision-makers are only consumers of knowledge, since they gather information about an external and fixed environment. In a strategic-like approach, however, they are also producers of knowledge as they are conceived as capable of affecting, through their reciprocal conjectures, one another's decision-set, rendering the environment more internal and liable to change. Hence, this qualification has been advanced with the purpose of outlining the epistemic assumptions that dwell in the models I consider here and, in accordance, proposing a more complex notion of rationality.

In this way we can make sense of the rather intriguing cases where the agent wants to design the result (as if in the intentional model) and fails altogether (as if in the causal model) for two rather distinct **epistemic** considerations (in the one case because the actor sees his environment as external and fixed, in the other because he sees it as endogenous and mutable. These are the paradoxical worlds in cells II (e.g., social choice theory) and IV (public-goods literature). We may then take an interest in following the ways-out conceived by proponents of these images. In particular, I would suggest here that models in cell II (parametric-public) tend to design a **moral** conception of individuals on the way to a solution (I am thinking of Rawls and Harsanyi's normative theories), whereas models in cell IV (strategic-public) tend to fail in the absence of this moral premise (e.g., public-goods literature).

More generally, models in the bottom half of the matrix, cells III (strategic-non-public) and IV (strategic-public), tend to find more stable deductive solutions in their parametric counterparts, cells I and II. For example, standard game theory tends to develop much in the

same line as a general equilibrium framework on the way to solution concepts,[8] and standard public goods literature has often found a way out in its implicit injection of some cooperative willingness (a moral premise) on behalf of its rational agents,[9] which is better worked out in normative approaches (in cell II). Apparently, then, the parametric field exerts an irresistible magnetic attraction on its strategic counterpart, while the distinction between the aspects of consumption and production of knowledge is somehow impoverished:



Recall that these four cells represent different nuances of the same basic idea that the social order results from the rational choices of individuals, the nuances having to do with the scope of the ends pursued and the description of the environment as more or less 'rationality-friendly' in terms of the knowledge available. The expected result is that rational individuals can coordinate their efforts relying on rationality alone; it is also claimed that this coordination is efficient and just or 'well ordered'. Cells III and IV will be discussed in chapters 3 (on game theory) and 2 (on the public-goods literature), in Part I. Cell II, meanwhile, will be the subject of Part II (on Arrow's social choice theorem and normative theories). Taken together, cells IV, III and II stand respectively for the efficiency, equilibrium and justice 'normative' attributes of a conception of social order and peculiar knowledge assumptions.

Now, a non-reductionist although to a certain extent still individualist thinking depicts

[8] See Kirman, 1992. See also chapter 3.

[9] See chapter 2.

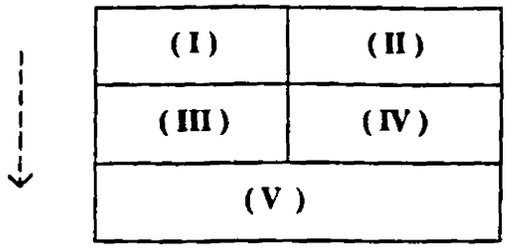
individuals as incomplete choosers who must have their incomplete rationality supplemented by evolving and irreducibly intersubjective multi-level rules on their way to building up the social world. In this case, the gap between intentions and actions is somehow filled by social forms of knowledge that cannot be fully articulated at the level of the individual consciousness and that are nonetheless acquired individually; indeed, even intentions are somehow conditioned by these individually acquired social forms. Let us call this view an **evolutionary** notion of rationality that may be illustrated by Hayek's social philosophy and is the subject of Part III. This account adopts a complex vision of the role of knowledge in our conception of the social world which emphasizes both its social and subjective aspects (as individuals are seen as producers of knowledge by the very act of consuming it). It also points to the limits of normative reasoning due to that complex picture. The incorporation of this new outlook expands my former matrix into the following:

MATRIX 2

R\E	non-public	public
parametric		**Part II** (Arrow's social choice theory; Rawls & Harsanyi's normative theories)
strategic	**Part I** (Game Theory)	**Part I** (Public Goods & Collective Action Literature)
evolutionary	**Part III** (Hayek's Social Philosophy)	

With the addition of this evolutionary perspective, I will assume that this field exerts an even stronger magnetic attraction on the strategic field than the parametric (though this can be only

partially illustrated here with the cases in cell II), and that it may also provide a better stand for the parametric field.[10]



Having thus characterized the landscape, the thesis discusses four cases out of the five depicted in matrix 2 (cells II, III, IV, and V) with a view to outlining their vision of the social world and, in particular, the degrees of **reductionism** and **normativity** involved in their 'individualisms'.[11] In this latter regard, I assume that in spite of current understanding the four cases I examine in the thesis are nonetheless normative to a certain crucial extent, and that this normativity is related to the way in which these theories deal with the knowledge issue.

One way of structuring the material is to move from the more reductionist and more idealized views to the lesser, or from the simpler to the more complex, in other words, from the 'one' to the 'many'. Another way involves the dramatization of the sequence where we may think of the Parts as addressing each other's weaknesses. This latter option is undertaken here mainly for rhetoric reasons, as I am not claiming that particular issues unanswered by a framework are settled by another, let alone that one approach succeeds in settling its predecessor in the sequence. What has interested me most is collecting, interpreting, and discussing well-accomplished arguments, within the general typology proposed in my modified matrix, regarding the rather 'hot' issues that dwell in the individualist tradition and which constitute the basic foundations of its view of social order: the normative and the epistemological. My ultimate purpose is to see the extent to which what these arguments achieve is affected by

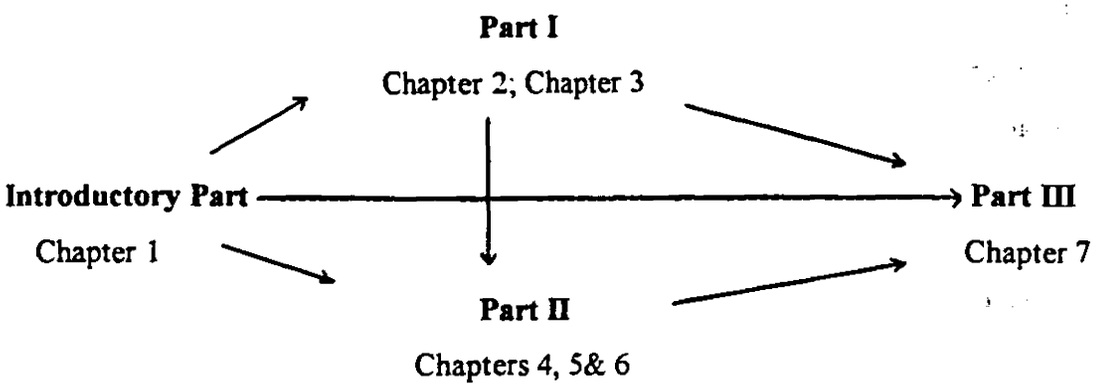
[10] Note that the numbers in the matrix still refer to the cells and not to the Parts of the thesis.

[11] Terms such as 'individuals',parametric', 'strategic', and 'evolutionary' rationality, and 'public' and 'non-public' purposes are to be given conceptual density as long as the authors whose ideas we are discussing provide the necessary elements to this end.

their epistemological stand.

To this end, the sequence adopted seems quite natural. So, I begin in chapter 1 by proposing a discussion of self-interest from the perspective of an eighteenth century European tradition, as the starting point of the set of theoretical and philosophical propositions that are examined in the thesis. It is an interpretation that downplays the explanatory and normative abilities of the behavioral assumption of self-interest. I proceed, in Part I, by discussing two theories that disregard the advice implicit in the tradition and assume the possibility of self-transcendence from a self-interest viewpoint ("Self-Interest and Beyond"). Self-interest can account for the public good of order in chapter 2 (on order as a public good) and can provide a basis for mutual coordination in chapter 3 (on order as strategic interaction). However, we find that a moral assumption is actually rather out of place in chapter 2, to the effect that the public good of order can only be provided if cooperative behavior is assumed on the part of the individuals which is at odds with the theories' starting point. Part II then follows in a natural way as it takes up moral arguments and assumes that the (well-ordered) social order has to rely not only on self-interest but also on a moral motivation of the individuals' ("Autonomy: Beyond Self-Interest"). In the same way, insofar as we have an epistemic issue in chapter 3 that nonetheless received only a rather light treatment, where the equilibrium condition depends on a postulated convergence of beliefs, Part III is proposed as an in depth approach to the knowledge conditions and their hopefully coordinating role in a conception of the social order. Finally, the epistemological discussion in Part III also addresses normative arguments raised in Part II as a certain view of knowledge, of which the argument in Part III is critical, is in fact crucial to the normative outlook taken up therein. The epistemological starting point in Part III also has effects on ethical arguments since it outlines the limited ability of reason to articulate our sense of justice. However, it is then argued that space for conscious and active intervention cannot be ruled out solely on epistemological grounds. The result is that the 'interacting individuals' version of the social world displayed in Part III proposes a non-reductionist understanding of it that still arguably encompasses horizons for hope. Moreover, this vision is more in line with chapter 1 where it is claimed that self-interest, or individual rationality for that matter, needs a context in order for it to mean anything.

A simplified scheme may depict the links between ideas very briefly:



Of course, I realize that to advance this scheme as a thorough integration proposal would require a different treatment from the one I have undertaken here. My main aim is to propose an **interpretation** of a set of theories in terms of the particular ways in which they view the social order as resulting from the interaction between its individual members, keeping an interest in the reductionism and normative purposes involved. Having spelled out their 'individualisms' as well as the epistemological and normative presuppositions, an effort of theoretical integration might then follow as subject for future research.

WHAT IS THE VALUE OF SELF-INTEREST?

What proposition is there respecting human nature which is absolutely and universally true? We know of only one: and that is not only true but identical; that men always act from self-interest. (...) But in fact, when explained, it means only that men, if they can, will do as they choose. (...) But we gain nothing by knowing this (...) In fact, this principle is just as recondite, and just as important, as the great truth, that whatever is, is. (...) [I]t is (...) idle to attribute any importance to a proposition, which, when interpreted, means only that a man had rather do what he had rather do.[1]

1. Introduction

In approaching self-interest one is easily led to seek its roots in the history of ideas. One reason for this is that it seems unhelpful to begin with an analytical scrutiny of the concept in order to find what self-interest should be. Another reason is that taking the concept at face value, and accepting its 'realistic' appeal, does not explain its many puzzling aspects. In order to avoid engaging directly in the deciphering of these conundrums, we shall pursue a somewhat crooked path, and follow the development of this notion in political arguments of the XVIIth and XVIIIth centuries in connection with some benefits it was expected to bring to social and political interactions at the time, if not to social and political order *tout court*.

Accordingly, it seems to be well established that the notion of self-interest came to attract a considerable attention as an important piece of political argument around the mid-seventeenth century. It is also generally considered that the concept reached its heyday in the eighteenth century, in that not only was it widely used to explain phenomena but its possibly virtuous by-products were also appreciated and even hailed.

Less remarked in its subsequent use, however, a third element besides the above explanatory and normative aspects has surfaced in my reading. This is that, though an increasingly powerful category, self-interest was forcefully taken as a **negative** or residue idea - a defensive claim for the existence of something (or someone) else beside the State or the love of God that should be taken into account as a new element within political arrangements and arguments. The 'self's-eye view' appeared as a rather non-place, a space open to political dispute.

---

[1] Macaulay,1978:124/125.

Of course this latter aspect very much qualified what could be expected merely on the basis of the two other claims, namely, the explanatory and the normative claims, in that the expected explanatory and normative virtues of the category 'self-interest' were somewhat constrained by its negative trait, and for good reasons. We may anticipate two of these reasons: that self-interest emerged as a mere *cogito* - the epoch attempts to build up a new standpoint from the disagreeable alternatives; and, that whenever taken too seriously it would reveal symptoms of endogenous decay.

This chapter is divided into two additional sections and a brief final comment. In section 2, it is argued that the category self-interest was taken up, in the XVIIth and XVIIIth centuries in the context of the debate concerning the content of the public interest. In particular, it was then suggested that the public interest should be somehow compounded of the many interests of the people, but also that the plurality of the interests should be somehow constrained by what was acceptable for all. In subsection 2.1, I outline the politically unstable state of the world that first gave rise to conjectures concerning the political benefits that might result from the inclusion of the interests of the many subjects. Subsection 2.2 presents a number of diagnoses at the time that identified factionalism as the source of political instability. In subsections 2.3, 2.4 and 2.5 self-interest is discussed with respect to its ability to overcome political instability, while subsection 2.6 presents an argument against altruism.

In section 3, I introduce a number of the ways of integrating private and public interests which were brought about in the eighteenth century. From the arguments given in the preceding section, follows the idea that the public interest should express a compatibility of the many private interests; a question then arises as to what forms this compatibility took in eighteenth century arguments. I have suggested that three models of compatibility emerged: that conflicting interests were seen in some pieces of literature as the efficient causes of public compatibility; that though conflicting, the many interests were still expected to be compatible; and the somewhat hybrid belief that though the public interest is materially caused by the competition of the many interests, there still remains some undesirable effects from this form of association, the many interests causing both common goods and common 'bads'. (It is argued in subsection 3.5 that the undesirable effects of a society moved by the self-interest of its members might be overcome by

the Lockean idea that the General Good should somehow enter into the very constitution of the ends of the individuals, shaping their ends as citizens. But this is another, non self-interest based, story.)

2. The State of the World, The Causes, and The Solution

2.1. 'Masterless world'

There is some literary evidence that Europe in the seventeenth century was seen by many at the time as a 'masterless world' which had come to that state of affairs due to a peculiar constitution of the public space, as an arena, that is, of the power of the privileged.[2] The public sphere thus depicted was regarded by many as unacceptably unstable. A revision of the extension of the polity was in order, or, at least, since the established model was collapsing, there was room for such a claim to revision.

Gunn (1969) reports on this interesting public debate which took place in England in the mid-seventeenth century on the meaning of 'public interest'. Much of the rationale of supporters of the Parliament was that this latter would better represent genuine public interest to the extent that it opposed the King's interest with a variety of interests of the **many subjects** of which it was compounded. The common condition of subjection and the multiplicity of interests (not easily reduced to one head of private interest) would prevent any privileged view from prevailing.

Public interest was interpreted, in this version, as being (loosely) compatible with the many interests of the subjects, and, as it were, as opposed to the one held by the King. To be sure, these many interests had different interpretations according to the various political principles maintained by their leveller, radical, or democrat supporters. However, they all had in common an eagerness to question the identity between the public interest and the interest of the King or, as it used to be paraphrased at the time, the King's duty to his people.

Much of this opposition, Gunn argues, was linked with a conviction that one should look elsewhere for the grounds to justify a political order beyond law and historical precedent. One

---

[2] See Gunn (1969), Mansbridge (1990) and Hirschman (1977).

such justification was provided by the doctrine of the natural right and its peculiar identification of reason with self-preservation.

The path from historical precedent or religious authority to the doctrine of natural rights thus interpreted can be easily followed through the literature that has looked into the different forms taken by the doctrines of social contract, as in Gough (1957). In particular, Mansbridge (1990) makes the rather sharp point that social contract doctrines before Hobbes used to describe the so-called state of nature as a state of war of some against others, among different factions, but not as a generalized state of nearly open conflict, of all against all. After Hobbes, however, opinions changed.

The political problem within the terms envisaged by the traditional way of thinking, with their stress on partisan struggle, seemed to have reached a dead-end. We may conjecture, then, that alternative arguments were initially motivated by the need for a novel description of the political problem. In particular, distinct descriptions were welcome which would hopefully discredit the belief that factionalism was a **natural** ground for a political order, since by natural was meant 'doubtless' or 'self-evident', hence, outside the reach of critical scrutiny.

We may be tempted to conclude that the instability and subsequent suspicion surrounding the naturalism of prevalent notions of public-interest appear to have inspired a re-description of the political problem as a somewhat non-natural one. In particular, suitable reasoning, it was sometimes suggested, might help us to decompose factions into smaller units, and, by breaking their bonds, other links could be substituted which would be more congenial to a stable political order.

One such argument, as is well known, describes the political problem as that of overcoming a 'natural' state of equality among individuals, a state of universal conflict calling for an overruling authority. Or, to put it differently, a vision where (overall) conflict not hierarchy is 'natural'. The roots of this 'naturalness', though, are to be found in the demands, logical or otherwise, imposed by this thought experiment.

2.2. Factions and Trump: even if 'the way be broad enough'

Again, why should one postulate that individuals, not parties or factions, are the units of the 'commonwealth'?

In his essay "Of Parties in General", Hume argues that a faction displays a sense of cohesion that competes with that of nation, a cohesion which surfaces from engendering a difference, practically an 'animosity', among equals:

> the influence of faction is directly contrary to that of laws. Factions subvert government, render laws impotent, and beget the fiercest animosities among men of the same nation, who ought to give mutual assistance and protection to each other. (Hume,1994:34)

Hume identifies three kinds of factions, originating from interest, affection (an 'imaginary' interest), or principle. Of these, the worst is that stemming from principle because it is the least accountable in reasonable terms and, accordingly, potentially more disruptive in its political effects. The rather shocking characteristic of this latter is that it may engender a conflict where there is no scarcity, thus revealing a disturbing and basic aspect of human nature:

> Two men, travelling on the highway, the one east, the other west, can easily pass each other, if the way be broad enough: But two men, reasoning upon opposite principles of religion, cannot so easily pass, without shocking; though one should think, that the way were also, in that case, sufficiently broad, and that each might proceed, without interruption, in his own course. But such is the nature of the human mind, that it always lays hold on every mind that approaches it; and as it is wonderfully fortified by an unanimity of sentiments, so is it shocked and disturbed by any contrariety. (ibidem:36/37)

This contention by Hume suggests the basic, though no less problematic role of 'opinion' in social interaction, as he considers opinion to govern people's actions. When principles held by people have an expression in different forms of actions, he argues, the relation is straightforward and we may understand the ensuing conflict. But when different principles do not have a necessary extension in conflicting practices, referring rather (as in Hume's time) to theological dogma or a proposition concerning the divine character of royalty, it reveals something else about human nature. It indicates that we are looking to convince others of our ideas, to take hold on their minds. Thus, the connection between opinion and action becomes importantly enough

problematic. We might foster conflict where there need be none. Furthermore, it is hard to see how a legislator could overcome this tendency.

A dangerous potential was also ascribed to opinions by Hobbes, as is well known. The confusion caused by opinion, according to Hobbes, has its expression in speech itself, the faculty of naming things and creatures:

> The names of such things as affect us, that is, which please, and displease us, because all men be not alike affected with the same thing, nor the same man at all times, are in the common discourses of men, of **inconstant** signification. For seeing all names imposed to signifie our conceptions; and all our affections are but conceptions; when we conceive the same things differently, we can hardly avoid different naming of them. For though the nature of that we conceive, be the same; yet the diversity of our reception of it, in respect of different constitutions of body, and prejudices of opinion, gives everything a tincture of our different passions. (Hobbes,1968:109/110)

Language is contaminated by opinions and these, as it were, are coloured red by passions. Where are we to find the 'true grounds for ratiocination', for in ratiocinating we merely resort to words and their consequences?

> Not but that Reason it selfe is always Right Reason, as well as Arithmetique is a certain and infallible Art: But no one mans Reason, nor the Reason of any one number of men, makes the certaintie; no more than an account is therefore well cast up, because a great many men have unanimously approved it. And therfore, as when there is a controversy in an account, the parties must by their own accord, set up for right Reason, the Reason of some Arbitrator, or Judge, to whose sentence they will both stand, or their controversies must either come to blowes, or be undecided, for want of a right Reason constituted by Nature; so is it also on all debates of what kind soever: And when men that think themselves wiser than all others, clamor and demand right Reason for judge; yet seek no more, but that things should be determined, by no other mens reason but their own, it is as intolerable in the society of men, as it is in play after trump is turned, to use for trump on every occasion, that suite whereof they have most in their hand. For they do nothing els, that will have every of their passions, as it comes to bear sway in them, to be taken for right Reason, and that in their own controversies: bewraying their want of right Reason, by the claim they lay to it. (ibidem:111/112)

And afterwards, he concludes that:

> The light of humane mind is Perspicuous Words, but by exact definitions first snuffed, and purged from ambiguity; **Reason** is the **pace**; Encrease of **Science**, the way; and the Benefit of man-kind, the **end**. And, on the contrary, Metaphors, and senselesse and ambiguous words are like *ignes fatui*; and reasoning upon them is wandering amongst innumerable absurdities; and their end, contention, and sedition, or contempt. (ibidem:116/117)

Hume and Hobbes both assume that opinion is the problematic source of conflict, and at the same time Hobbes suggests that it is the material we should try to improve. We should erect the right Reason in the person of an arbitrator that must not advance one particular opinion as the yardstick. And it is clear that this viewpoint is to be artificially constructed. The meaning of politics as an art surfaces here, and one is tempted to compare this sense of art, *pace* Hobbes, with the Greek *techné*, which is a kind of knowledge that deals with material that contains contrary forces, as opinions might be potentially disaggregating or aggregative (whether or not properly shaped by political argument).

As a side point, Hobbes' belief in the arbitrary origin of civil laws is usually taken to mean that according to him, any law is better than none. But it is also possible to interpret this latter contention, in view of the preceding reasoning, as meaning that no specific law is **demonstrably** better than any other, though one may have preferences among the possible range of laws, and Hobbes himself seemed to have spelled out his own preference for a specific form of government. In my judgement, though, his emphasis is rather on the logical necessity of an artificial procedure, a system of references by which the conflict of opinions might be reasonably reduced, a procedure he deems to be quite demonstrable, whereas a reasonable defense of a specific set of laws would require us to face the question of 'trump'.

2.3. Self-preservation

It is quite clear that to construct the arbitrator's viewpoint, or the place of the public interest, Hobbes would need to clarify its bearing on people's interests, for although it cannot stem from one 'mans Reason', nor from 'one group of mans Reason', not even from a great many men's approval, it must have a hold in people's interests. We should look for something universal, so to speak, above the parts, that can be figured out by anyone. The natural law which obliges everyone

to look after their own preservation might yield a proper understanding of a situation of extreme conflict as well as a possible way out, Hobbes argues.

One's interest, therefore, could be safely constructed so as to envisage in some sense both civil peace and self-preservation. In this way, civil laws were thought necessary to convey a reasonably acceptable meaning to self-preservation, this latter as somehow committed to peace. Conversely, self-preservation would impose a considerable restraint on the laws themselves, where public interest was understood as the commitment to preserve people's lives.

In support of this view it is worth recalling some of Hobbes anthropology . Hobbes populates the state of nature with people whose private concerns lead to overall conflict. What is implied in these concerns? Their exact content initially deserved a rather extended taxonomic treatment by Hobbes, as when he came to list the passions in the 'Leviathan'. Thus, he seems to suggest, we should identify our irrational impulses and some of their clearer consequences before we come to define our interests. So, the first step is to get a map of the *selva oscura*. Then, he goes on to investigate our natural duties to ourselves, the so-called natural laws. But, then, we are still in a somewhat muddled territory, for our interests seem to shift from the natural and to some extent harmless drive for self-preservation to the drive for power over others. This uncertainty is sufficient to instill a broader instability for even the internal boundaries of self-preservation, he seems to suggest, may be not so clearly demarcated. Self-interest has a potentially dangerous strategic component which may make the death of one's neighbor the safest way to attend to one's natural duties to oneself. The suggestion, which we are hardly compelled to resist, is that we should somehow create the circumstances where self-preservation is compatible with peace, that is, with a common understanding of what self-preservation is.

Hence, the message of (the laws of) nature is not so straightforwardly comprehensible; we need a translation in terms of civil laws. These, in turn, would constrain the content of self-interest, through clear obligations, thus complementing the lack of information with respect to our real interests.

We may then see that to a certain extent the definition of the public interest enters into the definition of self-interest in that the former raises barriers to the latter's extension. Besides, these barriers appear to be somewhat constitutive of the meaning of self-interest as well, as when we talk about 'real' interests. That is, again, this view 'constructs' the self-interest as that portion of human motivation which is congenial to civil peace.

In support of this view, it is interesting to refer, again, to Gunn's report of English public debate at the time. During the mid-seventeenth century debate there were some who maintained that the public interest was not the King's version of it but the common interests of the rich, or their property interests, for as proprietors they were deemed to be more concerned with the nation's affairs. Subsequently, during the debate, property rights took the rather expanded form of private rights, and the common interests were thus interpreted as private rights. These rights, in turn, were defended as at least compatible with peace and order, in a typical manoeuvre to minimize 'the amount of tension between particular interests and the general interest'.[3]

In the same spirit, the demand for private rights was interpreted by some as a demand for the expansion of the realm of individual liberty. Concerning the definition of public interest Gunn writes:

> Gradually there was a shift in emphasis from recounting a king's duties to his people to defending people's right to look to their own interests. (Gunn,1969:16)

This quest for a degree of privacy in the understanding of the public was expressed in the fact that political arguments, according to Gunn, were moving 'from grounds of law and historical precedent to grounds of reason and natural right':

> Some of the earliest Parliamentmen had chosen to plead from the laws of necessity and nature, and especially that of self-preservation. (ibidem:17)

A sort of unanimity arose with regard to one meaning of private interest that should be rejected altogether: that no man should be judge of his own case. Here we cannot help recalling Hobbes' 'trump' and Locke's concerns on the matter. Still, following Gunn, this claim was taken over by

---

[3] Cf. Gunn, op. cit.

the Parliamentmen to the effect that Parliament should be either an impartial arbiter between the King and the people, or whenever it could not possibly claim impartiality, it should decide the case not in its own interests but in those of the people.[4]

Although there seemed to be unanimous rejection of partiality in the above sense, an unanimous positive standpoint was still missing. Actually, the concern with self-preservation, as Gunn points out, only moved the debate one step further, for the meaning of self-preservation, and what is required to strive for it, was far from clear. People had fought precisely for different meanings of it. Some maintained that self-preservation meant property rights, others like the democrats of the time disagreed. It was claimed, for instance, that 'taking away a man's property was the effective equivalent of taking away his life'.[5] And, although the general Leveller viewpoint attempted to emphasize that 'in order to give the public interest some content' we should care for particular interests, some objected that these interests interpreted as property rights were but a measure of official oppression, or, as John Warr, a seventeenth century democrat, believed

> The people's good had always been the professed of government... The people had to be given an opportunity to express their interest, a good as understood by themselves. (ibidem:29)

And here, Gunn argues, we unravel a concern with the enlargement of the area of private freedom which was far more demanding than the claims for the preservation of the rights of private men, as these had been previously understood (and, thereby, the commonwealth).

The interests of the many, we may conjecture, could be either the **common** interests they have, as in their expression under the property rights movement, or as potentially more conflictual with greater diversity, as implicit in the democrats' claims for a more complex (and less aggregative) concept of the public.

> Like later democrats, those of the seventeenth century tended to move from the position that common men best knew their own interests to the position that they might then understand the public interest. They were not concerned with all of a

---

[4] Cf. Gunn, op. cit., p.18.

[5] Cf. Gunn, op.cit., p.20.

> man's private interests, just those that most closely impinged upon public policy. Nor were they very explicit about the interests of single individuals, preferring to ascribe rationality to the public. However, they were obviously implying certain things about individual men. It happened that the individual interests of concern were rights valuable to all, and, indeed, insecure unless all shared them. Securing life and property was the *sine qua non* of further individual fulfillment. The connection between the public interest and those pressing particular interests that could be satisfied by political means was extremely close. (Gunn,1969:30/31)

We come to see that, in this debate, a somewhat bewildered identity of interests with public and private, was being worked out. The rejection of a trump or a partisan perspective was carefully made explicit, as well as a loose interpretation of the public as somehow assuring the *sine qua non* of individual fulfillment however it was to be understood.

One particular instance of an unacceptable view of conflict is given by a change in the connotation of the word 'party' in the mid-seventeenth century and the relation of that change with a similar modification in the idea of conflict itself:

> In the mid-seventeenth century, Parliament, which had almost all its decisions by consensus, began making most decisions by majority vote.(...) Political parties began to develop into ongoing organizations that remained in conflict from one election to another, and the word **party** began to lose its unsavory connotation of a faction opposed to the common good. (Mansbridge,1990:6)

The demand for cohesiveness began to fade from the immediate agenda and to move to a farther and more permanent (however imprecise) background.

Thus, within a sound political environment, parties, and conflict between them, were no longer viewed with reproach. Parties, in the old sense of the term, were to give up their blind (and lexicographical, so to speak) allegiance to some revealed conception of the common good.

In conclusion, having resisted the received meaning of public interest, the epoch has come to make a case for self-interest and to rehabilitate it from its pre-modern exile. At the time self-interest must have appeared a somewhat refreshing idea, once it had been purged of all partisan connotations.

2.4. 'Read Thy Self': Self-interest as a *Cogito*

At this point, I venture, two related but different thoughts spring from the image of a 'masterless world'. The first is a sort of anti-naturalistic outlook concerning the political realm, to the extent that the prevailing political order appeared unstable. The question arises, What is 'natural' in the political domain? Certainly not the traditional way of doing nor the traditional way of justifying what has been done.[6]

Images of the political body as an **artificial** construct abounded from the seventeenth century on, the most famous was Hobbes' which somehow echoed and radicalized an earlier Machiavellian insight into statecraft. Much of the idea behind social contract theories, according to Gough, is that any political society had to be understood in terms of a kind of **arrangement** between individuals.

The other thought, which follows the preceding, refers to a reflection on the subject who is criticizing the received view on the foundations of political order, a subject that is born as a sort of critic. The idea is that self-interest is built up as the motivation of this new political subject (albeit a residue of other motivations) and as a sort of *cogito* of the political experiment. We may be suspicious about the trustworthiness of all our motivations, yet the doubt is there. Is this doubt grounds enough to support a reliable political space? What could possibly be a 'Self's-eye view', in contrast to (or in the abstraction of) God- and King's- eye views? These are questions, I presume, that were on the XVIIth and XVIIIth century agenda.

Thus, recall the Hobbesian 'Read thy self', which is meant to

> teach us, that for the similitude of the thoughts, and Passions of one man, to the thoughts and Passions of another, whosoever looketh into himself, and considereth what he doth, when he does **think**, **opine**, **reason**, **hope**, and **fear** &c, and upon what grounds; he shall thereby read and know, what are the thoughts, and Passions of all men, upon like occasions. (Hobbes,1968:83)

---

[6] This anti-naturalistic outlook that seems to have emerged in the seventeenth century could arguably be taken as an act of 'self-consciousness', following the facts pointed out both by Mansbridge (of religious authority faltering in the face of actual conflict) and Gunn (of historical precedent failing to provide argument enough to prevent resistance).

Hobbes invites his reader to engage in critical inquiry, which involves introspection and comparison. In his suggestion, such a reading of ourselves might enable us to define a common path. An important part of this inquiry is then turned into a concern with the driving forces that motivate individuals, and the corollary question of how is it possible that anyone can effect a genuine detour from his particular experience to the more general grounds involved in mankind? This is especially important given that he who is to govern a whole Nation 'must read in himself, not this, or that particular man; but Man-kind'.[7]

There is certainly something abstract in this critical scrutiny, in the mode of work which is set in motion typically in political philosophical reasoning. In political philosophy, says Rawls,[8] the work of abstraction is set in motion by deep political conflicts. The deeper the conflicts the higher one has to elevate one's thoughts in order to achieve a somewhat reasonable common ground among the contending parties.

My suggestion of self-interest as a political **cogito**, encompasses some of Rawls' insight regarding the conflictual nature of the subject, and also the necessity of finding a more abstract perspective, but concludes with a persisting (though enlightened) doubt. In the 'career' of self-interest, somewhat artificially traced here, I identify a search for a more definite content whose conclusion is still a doubt.

Moreover, in this sense, the notion of self-interest cannot claim precedence in the search for a meaningful sense of public interest. I shall take the view that the category self-interest emerged (or re-emerged in its modern clothes) from a critical inquiry into the meaning of public interest and became so intertwined with the definition of public interest that it gave rise to the ensuing confusion, that still persists concerning the normative character or optimal qualities of selfishness, and which has nearly nothing to do with its canonical Smithian (let alone Hobbesian and Humean) form.

---

[7] Cf. Hobbes, op.cit., p.83.

[8] Rawls, 1993(a), p.44.

2.5. Approaching the meaning of self-interest: a limited claim for universality

Some meanings of interest in the seventeenth century, Gunn reports, were 'concernment' and 'importance' to oneself. There was also the meaning of interest as the nature or the normal function of a thing.

> People occasionally referred to the interest of the inanimate objects. It was one way of naturalizing the term, by equating it with that law of nature commonly thought to be reflected in the activities and nature of all things. [Charles] Herle showed his acceptance of this usage when he described interest as 'the center of everything's safety' and noted that one might then speak of a stone's interest in obeying gravity and staying on the ground. (ibidem:43)

Less physically (or in a more moral fashion), interest, according to Hirschman (1977), was worked out in arguments concerning the possibility of a political order of the early seventeenth century, as an alternative to the disagreeable options of passions and reason. Passions were viewed as destructive and reason as ineffectual. Interest, in turn, was claimed to gather the benefits of both motivations, namely, the element of reflection contained in reason as well as the strength normally associated with the passions.

This rather normative character of the notion of interest soon gained an illegitimate heir, when arguments advanced at the end of the seventeenth and throughout the eighteenth century exalted the explanatory qualities of the category, as in Helvetius's famous physical analogy of the laws of motion in the physical world and laws of interest in the moral universe.

The distinction between uses and abuses of the notion of interest is better felt through the contrast between the following two maxims: the early-seventeenth century dictum that 'Interest will not lie', and thus was hoped to constitute a more reliable basis for a political order than traditional sources, and 'Interest Governs the world', a late seventeenth century maxim and a comparatively explanatory claim that supersedes the artificial, normative, character of the category, as Hirschman notes.

In any case, predictability and constancy, Hirschman goes on to remark, were nicely neutral qualities usually expected from self-interested action in the seventeenth and eighteenth centuries.

In the same vein, Holmes (1990) contends that self-interest appeared attractive when compared to the disagreeable alternatives, examples of which abound in the writings of Hume, Smith, and others, such as self- and socially- destructive behavior, and that its attractiveness was to be found not so much in its explanatory reach as in its expected normative qualities. It was understood as a set of comparatively less noxious forces that drive human action, thus compatible with social interaction.

Holmes invokes, for that matter, the Humean distinction between self-interest and factionalism. This latter was claimed by Hume to be more basic, as a sort of inborn tendency to identify emotionally with an exclusive group for the deeper reason of an even more basic human feature, namely insecurity. Shared opinions hold groups together, but may lead individuals astray (and society as well). Interests and all human affairs are, in the end, governed by opinions, as a consequence, people often fail to grasp their real private advantage or to act upon it when they do grasp it. (We may wonder whether this is why Smith preferred to substitute trade for persuasion, when he elaborated what he felt to be a basic human inclination, namely, the 'commerce' of ideas.)

Again, the suggestion of a fluctuating content of self-interest implies that this might be supported by deeper-rooted opinions, as those stemming from proper political argument. In particular, we may conjecture, in a somewhat Humean spirit, that whenever people follow well-established conventions they are acting from well-grounded opinions.

In any case, Holmes maintains, a fundamental reason for the intellectual success of the category of self-interest lies in its potentially egalitarian and democratic connotation, against the backdrop of the alternatives. Thus, universal self-interest makes people discover conflicts of interests everywhere, particularly between the interests of citizens and those of the wielders of power. From this postulate, people are entitled to a certain scepticism with respect to any self-proclaimed manifestation of 'public interest'. Smith, Holmes goes on to argue, used the postulate of self-interest to criticize the idea that harmony was a natural product of hierarchy and subordination, and supposed that harmony could be the (imperfect) result of conflict, instead.

In any case, Holmes and Hirschman stress, self-interest, in the good tradition of Hume, is but a prudential maxim, something one predicates of one another, as a vicarious presumption, we might still say, that may be made to support a stable political arrangement.

We come up then, after Hirschman and Holmes's remarks, with a stress on the normative content of self-interest, and normative in the rather weak sense of a category that is shaped as a restriction to other comparatively more deleterious motives to human action. Thus, in a somewhat surprising conclusion, 'self-interest' appears to have been born as a constraint!

Another course, taken by Hirschman (1986), clarifies a little further what it is that is restrictive about self-interest.

He further pursues the roots of the interest category in discourses related to statecraft as well as individual behavior, only to find Machiavelli and Mandeville at the beginning of everything and he performs a second turn on the self-interest screw. This time, we learn that the category in fact evolved from euphemism, in its sixteenth and seventeenth centuries forms, to tautology, in its eighteenth century and later expressions.

So, according to Hirschman, the concept of interest has evolved with time from euphemism to tautology. As a euphemism it used to hide, and in this sense to restrain, the real forces or driving mechanism behind either the sovereign's acts (as in Machiavelli) or individuals' behavior (as in Mandeville), making them work for the public. Interest, in Hirschman's view, was, in Machiavelli's discourse, an euphemism for cruelty, mendacity and treason, much in the same way as avarice and love of lucre, in Mandeville's writings.

It was fundamentally the aspect of rational calculation attached to interest, in its euphemistic function, Hirschman argues, which was given emphasis under its early expression in the XVIth and XVIIth centuries. Interest-propelled action meant the action that had undergone a calculation of costs and benefits, and this would mean a prior restriction on the range of possible actions.

The widespread use of the category was followed, according to Hirschman, by a feeling of mistrust toward activities that were aimed at achieving the public interest directly, as in Adam Smith. Smith declared, for that matter, that he 'had never known much good done by those who affected to trade for the public good'.[9]

In this connection, it is worth noting, again, that the normative sense involved here is quite negative, that is, it refers to expected widespread benefits from private actions or interests stemming from **restrictions** interest would impose on the range of possible actions rather than from a straightforward **identity** between the private and the public spheres. As a corollary, it also refers to the negative judgement attached to a non-mediate connection between the private and the public standpoints.

However, a thorough 'Invisible Hand Doctrine' was worked out, encompassing a measure of 'realism' in the consideration of human affairs as well as 'an attempt to prove that it is possible to achieve a workable and progressive social order with these highly imperfect subjects, and, as it were, behind their backs'.[10] This doctrine of identity of the interests of a part with those of the whole, with its seemingly forceful elements of 'realism' and 'alchemy', took the fairly vulnerable form of a paradox, argues Hirschman.

As Hirschman points out, although the 'Invisible Hand Doctrine' became highly fashionable in eighteenth century arguments, as the paramount form of the doctrine of the identity of interests, it also began to be attacked by those who felt there were other springs for action, such as sympathy, generosity and so on, and those who began to lament 'the world we have lost'

On the other hand, it is worth recalling that a substantial support to the doctrine of self-interest was given by utilitarianism. In particular, utilitarianism laid down a strong defense against the objections to the doctrine of interests raised by those who lamented its failure in expressing the multiplicity of human motivation. This defense greatly contributed to the form of tautology self-

---

[9] Cited in Hirschman, op.cit., p.40.

[10] Cf. Hirschman, op.cit., p.45.

interest has taken since the XVIIIth century onwards. This is the second element of a doctrine of self-interest Hirschman addresses.

As a tautology, interest was taken to cover any area of human aspiration and deed, and in this sense it would not impose any constraint on choices or motives as the difference between a passion of whatever kind and an interest had virtually disappeared:

> In the end, interest stood behind anything people do or wish to do, and to explain human action by interest thus did turn into the vacuous tautology denounced by Macaulay. (Hirschman,1986:50)

Moreover, it is worth mentioning the fact reported by Holmes (1990) that Mill had thought of the difference between self- and public- interest as one of degree, rather than of kind. An argument, it goes without saying, that leaves room for the aggregative conception of the public interest typically entertained in utilitarian arguments.

On closer inspection, the shift from euphemism to tautology in interest-based arguments represented a difference in emphasis in the two elements that Hirschman identifies as constitutive of the notion of self-interest, namely, rational calculation and self-centeredness. The implicit contention by Hirschman is that in as much as self-centeredness gains prominence in matters of 'self-interest', as in the utilitarian argument, a real restriction on behavior is missing and, as a consequence, any account of someone's actions in its terms will be but a truism.

As a matter of fact, the aspect of self-centeredness and the related problems had already received extended attention by Macaulay in his famous critique of Mill's Essay on Government. In his essay, Mill translated self-interest into the calculated search for a net balance between pleasures and pains. Macaulay then spelled out a warning against the idea of reducing every motivation to self-interest thus interpreted:

> What proposition is there respecting human nature which is absolutely and universally true? We know of only one: and that is not only true but identical; that men always act from self-interest. This truism the Utilitarians proclaim with as much pride as if it were new, and as much zeal as if it were important. But in fact, when explained, it means only that men, if they can, will do as they choose. When we see the actions of a man, we know with certainty what he thinks his interest to

> be. But it is impossible to reason with certainty from what we take to be his interests to his actions: one man goes without a dinner, that he may add a shilling to a hundred thousand pounds: another runs in debt to give balls and masquerades. One man cuts his father's throat to get possession of his old clothes; another hazards his own life to save that of an enemy. One man volunteers on a forlorn hope: another is drummed out of a regiment for cowardice. Each of these men has, no doubt, acted from self-interest. But we gain nothing by knowing this (...) In fact, this principle is just as recondite, and just as important, as the great truth, that whatever is, is. (...) [I]t is (...) idle to attribute any importance to a proposition, which, when interpreted, means only that a man had rather do what he had rather do. (Macaulay,1978:124/125)

What does self-interest (deductively) impose? Macaulay responds that we are unable to determine this.

But, what if the meaning of self-interest be somehow restricted?

Then, Macaulay replies, the 'doctrine that men always act from self-interest', though ceasing to be identical, ceases also to be true.[11]

I presume that by this statement he meant, in modern terminology, that the doctrine could be falsified, and, in this sense, could lose its deductive pretence. In particular, Macaulay believed that another motivation, the desire to be loved, should be put on a par with self-interest, in that it conceivably might also raise regulative barriers to human action.

As a conclusion, it seems that against the logic of parties or factions, or passions or ineffectual Reason, self-interest intimates a logic of universality, be it in the form of predictability and constancy, or egalitarian and democratic potential. This logic nevertheless proves hard to translate into a precise calculus (the elements of which are a bit cloudy) and therefore seems to give rise to an explanatory account of human action. The element of self-centeredness is not a very reliable basis and should not be exaggerated lest tautological reasoning become widespread. The element of rational calculation, in turn, seems to be more promising (but only vaguely so), insofar as it

---

[11] Cf. Macaulay, op.cit., p.125.

stands for an ingredient of concern with one's future (and thereby, hopefully, a concern with other people) that is deemed to be present in a self-interested action.

2.6. Self-interest and altruism

Why the stress on 'self'?

I quote Hirschman (1986), for a natural reason:

> when action is supposed to be informed only by careful estimation of costs and benefits, with most weight necessarily given to those that are better known and more quantifiable, it tends to become self-referential by virtue of the simple fact that each person is best informed about his or her **own** desires, satisfactions, disappointments, and sufferings. (Hirschman,1986:36)

But, even this 'inner thing' might be informed by some kind of 'otherness' as conjectured by Pizzorno (1986), in situations, that is, when the identity of the self, and hence, its aims and sources of satisfaction, are given by its identification within a collective entity and without this reference, that particular self vanishes as such and becomes a different one. So then, one is invited to think of the self as anything but a natural entity. Thus, to take the self as a natural entity is not, arguably, a good way to address its priority. He seems to ask, in quite a broad sense, 'what is the meaning of 'self' without any broader references?'

Another reason might emerge from the contrast between self-interest and altruism. One might wonder why this other-regarding motivation was not given the same attention as self-interest. In this respect we quote Elster (1990):

> There is a sense ... in which self-interest is more fundamental than altruism. The state of nature, although a thought experiment, is a logically coherent situation. But we cannot coherently imagine a world in which everyone had exclusively altruistic motivations. The goal of the altruist is to provide others with an occasion for selfish pleasures - the pleasure of reading a book or drinking a bottle of wine one has received as a gift. If nobody had first-order, selfish pleasures, nobody could have higher-order, altruistic motives either... The point is just a logical one. If some are to be altruistic, others must be selfish, at least some of the time, but everybody could be selfish all the time. But we cannot conclude, neither in general nor on any given occasion, that selfishness is the more widespread motivation. (Elster,1990:45)

Thus, according to Elster, a principle of parsimony (rather than realism) could be invoked as a defense of the 'methodological priority' of self-interest. The logical issue is sound enough and, perhaps, provides a reason for the success of social-theoretical explanations relying on self-interest. If we had to choose just one motivation and the candidates were (a well-established definition of both) self-interest and altruism, self-interest is conceivably a more basic motivation in terms of a coherent thought experiment. Short of a well specified experiment, we may still dispute the priority of the 'self'.

Still, another reason for the relatively scant success of altruistic motivation as compared to self-interest is provided by Larmore (1987), and it refers to the more demanding requirement that altruism would fail to pass the Kantian 'universalizability' test. Except under the very unlikely conditions, the argument goes, of unanimity of opinions or coincidence of interests, the scenario of generalized benevolence would fail to deviate from the Humean circumstances of justice, for conflict would be a more plausible result of individuals' striving for the public interest. These efforts, Larmore contends, would translate into different and irreducibly conflicting conceptions of the 'good life'. This is actually one of the objections Kant raised against the principles of happiness, that they might allow for a variety of judgements as to what universal happiness imposes.

But, on reflection, is self-interest more likely to pass the test? Of course the answer depends, again, on how one defines self-interest. Thus, if we think of self-interest as vaguely as the demands self-preservation would impose in a lawless world, the universalizability test would be far from proved. But, again, Kant discusses the case where self-love (or private happiness, as he calls it) might be elevated into an **objective** (that is, universal) practical law. And this might make the 'happiness of others' rely only upon the universality of self-love and not on an 'additional external spring', such as a sympathetic disposition towards others:

> let the matter [of the maxim] be my own happiness. This (rule), if I attribute it to everyone (as, in fact, I may, in the case of every finite being), can become an **objective** practical law only if I include the happiness of others. Therefore, the law that we should promote the happiness of others does not arise from the assumption that this is an object of everyone's choice, but merely this, that the

> form of universality which reason requires as the condition of giving to a maxim of self-love the objective validity of a law is the principle that determines the will. (Kant,1978:304)

Nevertheless, practical contradictions are expected to arise, and, according to Kant, to indicate an independent origin of our moral judgements. For example, we would not take the services of anyone as a steward who, we were told, had taken his duty to his private happiness to the point of stealing. However, it is arguable that this restriction would be important only in an ideal community. Hence, self-interest might limitedly survive (with some normative loss) the universalizability test, under, that is, a somewhat restricted range of possible meanings, one of which could even harbor a weak form of 'altruism without inclination' (as universal private happiness).

And what about altruism? Under Larmore's interpretation altruism has the sense of action motivated by one's conception of the good life, that is to say, by one's idea of what is or would be better for everyone. Does altruism necessarily require that much? One may think of many cases where concern for the happiness of another might straightforwardly translate into attending to his uttered needs. To be sure, we may easily think of cases where we doubt he gives true voice to his interests, and still, feeling concern for that person, we may engage in a conjecture on what his true interests are. We may then achieve some principles that might guide our altruistic endeavors towards that person. Subsequently, we may feel that these principles conflict with the presumed one guiding the explicit utterances of the individual, and, the next step left to us would be to further pursue a criterion that would reduce the conflict, and so on. Let us suppose this conflict to persist. Our determination to promote the other person's happiness might eventually establish the kind of 'conflict without scarcity' Hume feared so much .

But is it still a matter of **altruism**? This altruist looks more like Ivan Karamazov and his general concern with humankind and absolute contempt for the man next to him. Let us take the less caricaturable case where the altruist's duty is something in between attending one's voiced needs and the search for a consistent action stemming from a fully articulated conception of the 'good life'. This altruism would be something that calls for construction in terms of as well as within a political argument. Under this form, altruistic motivation is less easily dismissed.

A subsequent question we would have to address relates to the 'autonomy' of altruism, an expression taken from Elster, and we might ask: What is the meaning of 'having an interest in other's happiness'? Is this reducible to self-interest? As for the latter question, the answer may be in the negative. One might think, for example, that Macaulay's claim for an equal priority to the desire to be loved to that already conceded to self-interest, stemmed from a belief of his that this desire is derivative of a more basic love for others, and therefore not reducible to a primarily selfish motivation. Hence, the fact that some of my desires are fulfilled by other would be a derivative of my overriding desire to so please the others as to let them please me, rather than as rooted in 'self-interest'. And a possible reply to the effect that this is but a form of the selfish desire to be accepted by others would only stress the point we have been making concerning the artificial (for that matter, arbitrary) character of the issue, for a question would naturally arise why these others are so important to us. This would but engender an infinite regress, as it now seems clear

So, we may justifiably ask, is this question interesting enough to dwell on? It may well be that all we need is a category that sufficiently separates the 'self' and the 'others'. The problem with altruism, in this sense, is that to postulate it is already to assume what we would want to know: can we have a concern with others? How 'many' others? Is this foundation enough for the 'public' interest?

It is thus arguable that to postulate a difference between the 'others' and the 'self's' standpoints would constitute the first step in a critical inquiry into the meaning of the 'public', that is, the interests of all. So, the rather minimalist view shall be taken that self-interest perhaps carries the advantage, at least in the interpretation presented here, of uncovering the two distinct elements in this inquest, namely the 'self' and the 'others'. As a category, it does not take for granted that a concern with others overlaps with the public space (it is actually an open attack on this idea) and requires for that matter that a distinction be spelled out. It also makes no claims about the real motivation behind people's action. It claims rather to be a locus of critical reflection, at least in the proposed interpretation of it as a political *cogito*.

It may well be that we do not know the real motivation behind our actions; then the next-best is to have a place from where to assess candidate motivations. Besides, altruism may not be a strong enough defense against the risk of factionalism.

3. Self- and public-interest intertwined

3.1. The 'identity of interests' issue

It has by now become quite clear that research surrounding the meaning of self-interest involves a reciprocal limitation: on the one hand, assigning a place for self-interest in political arguments indicates a desired limitation of the content of public interest; on the other hand, finding the limits of public interest helps to shape more finely the contours of self-interest (in that civil peace constrains the acceptable meanings of self-interest, to say the least).

Two such different thoughts regarding the regulatory principles supporting the civil laws are provided by what we call a 'theory of the hedges', of laws, that is, that in the words of Locke, 'hedges us in only from bogs and precipices', and Locke's theory of 'Laws of Justice'.

The easiest parallel here is that with negative and positive freedom, as suggested by Locke himself. A contrast, by the way, that gave rise to momentous reflections in modern political thought. In this study, I shall not pursue this distinction, though referring to it on occasion, for it does not seem to be very helpful regarding the immediate purposes here. I shall be mainly interested in uncovering some of the motivations behind these different senses of the law that have some bearing on self-interest, that is, focus shall be put on the underlying views of compatibility between the private and the public spheres.

Gunn specifies the frame of interest as far as statecraft is concerned around the seventeenth century, identifying the elements of it as well as some new areas into which it would be expanded:

> A variety of factors forced the political argument of these years into the relevant patterns. The tendency to conceive of the pressing problems as domestic, the need of relating the interests of governors and governed and the appreciation that the significant political units were interests, all contributed to the pattern. Circumstances caused writers to frame an understanding of the public interest

> colored by the requirement that it should be a condition understood by common men and realizable by their activities; less persuasively, they sometimes suggested that somehow a union of interests might encompass all antagonistic forces. (Gunn,1969:53)

The public interest, these arguments have suggested, might be reconstructed as the compatibility of private interests. Working on this suggestion, we discern, insofar as private interests may be conflicting, two alternative insights into the connection between public and private interests, one pointing to an interplay between the two which is compatible **though** conflicting and the other pointing to an interplay which is compatible **because** conflicting.

This point needs some clarification. The idea is that when one departs from a monolithic view of the public interest, as implied in hierarchy, then one has to cope with the somewhat difficult issue of conceiving how the many interests will be compatible with the common interests of all, encompassing of course the more likely cases where full coincidence of those interests is missing. Having in mind, then, that conflict is at least a corollary of the multiplicity of interests, one has to find an argument that accounts for compatibility-*cum*-conflict.

Two possibilities have arisen: that private interests, **though** conflicting, are compatible with the public, or else are compatible precisely **because** they are conflicting. In the former case, the task is to prove how competitive claims might be rendered compatible, or at the very least to explain why we should not expect competition to be disruptive; in the latter, though, one has to prove an identity between harmony and conflict, that is, one needs a causal connection, a view of competition as **the** very mode of harmony. In this latter case we might think, for instance, that the 'social cement', acquiring an economic meaning - like money, or the 'general equivalent' - relieves the environment from direct animosity, the shockingly concrete forms conflict might and has taken.

In any case, the present hypothes is that in its modern cradle, that is, from its seventeenth century rehabilitation and maturing through the eighteenth century, interest was not properly speaking given the prominence achieved in the full 'compatible because conflicting' argument, though a weaker form of it was attempted in the eighteenth century. But this weak 'compatible because conflicting' argument provided an excellent occasion for us to meet the weaknesses of the interest

'project' itself, for in exaggerating the benefits that could unfold from the workings of mere self-interest, however barely conceived, it succeeded in spelling out the dangers of any utopia based on it.

As is now clear, I shall not espouse Elie Halévy's typology straightforwardly, at least not without any further reflection. In particular, it is contended that the argument of a natural identity or fusion of interests which Halévy ascribes to Smith's Wealth of Nations deserves a substantial restatement.[12]

* * *

A paradigmatic form of compatibility was worked out in the seventeenth century in a rather complete social philosophical form through the social thought of the descendants of French Jansenism. It displayed a clear-cut separation between social order and progress on the one hand, and morality, on the other.

This compatibility took the sharp and shocking form of a paradox, whereby the lack of morality was made to work for social progress. However, the intellectual sources and motivations of the rhetorical form the argument took have recently emerged.[13] According to Horne (1978), it was mainly aimed at criticizing contemporary conservative religious and social thought.

The social philosophy of Jansenists, specially Pierre Nicole's, worked out the rehabilitation of self-interest. It expressed essentially a double message, the one that the imperfect creatures individuals are, often not well intentioned towards their neighbors, they nonetheless might be made to work, precisely as they are, in the common interests of all. The second message, placed at a deeper level than the former, was a rather skeptical admonition, namely that individuals' intentions are hard to uncover, not only to others but also to themselves. Social progress is not a matter of morality but of how we can make the most out of a bad lot. This is

---

[12] See Halévy, 1972.

[13] Cf. Horne, 1978.

not meant to dismiss morality altogether, for Jansenists believed that 'the state of doubt' is morally sound, and it is well ingrained in human beings precisely because of the presumption of transcendence that man harbors as well as the absence of any evidence of its existence.

> The knowledge of our humility makes us proud, and the knowledge of our pride makes us humble. We are strong when we know ourselves to be weak, and we are weak when we think ourselves strong. Thus this obscurity which prevents us from seeing whether we act from charity or amour propre, far from being detrimental to us, is salutary. (cited in Van Kley,1987:77, quoted from Pierre Nicole)

So, in the seventeenth century Jansenists had rehabilitated self-interest in the form of self-love, to its full potential.[14] Pierre Nicole's social thought, for example, grounded self-interest in the peculiar Jansenist cosmology: man is entitled to self-love in so far as he lacks a metaphysical, external reference, for God's support is inscrutable. A prosperous society is then considered to result as an unintentional outcome of man's efforts at erecting himself with reference to the next-best thing he has available, namely, others in the same environment and the things of **this** world. Accordingly, a flourishing society may be shown to be a consequence of three basic human desires: the fear of death, the love of lucre, and the desire to be loved. These constitute what Nicole calls 'enlightened self-love' which may take remarkably congruous forms in social intercourse:

> although nothing is more opposed to charity which relates everything to God than self-love (**amour propre**) which revolves entirely around the self, yet there is nothing more similar to the effects of charity than those of self-love. So closely does it follow the same paths that one could hardly do better in marking those to which charity should lead us, than to discover those actually taken by enlightened self-love. (cited in Van Kley,1987:73, from De la Charité et de l'amour propre)

The lack of a metaphysical reference, though, leaves an indelible mark: man is always in a state of uncertainty as to the true premises of his and others' acts.

> the principles of charity and **amour propre** are no more introspectively distinguishable than they are in their external effects. Behind 'formal and express reflections' flit 'transitory thoughts', 'confused ideas' as well as 'imperceptible movements of the heart', which escape conscious attention and render one aware

[14] A full report of Pierre Nicole's social thought is found in Van Kley (1987), and pieces of his argument are referred to in Horne's (1978) comments on the social thought of Mandeville.

> of the complexity of one's own motivations. Charity and amour propre are therefore often found mingled together in one's motivations 'without his being able to know for certain which of the two prevails.' (Van Kley,1987:77)

A congenial standpoint was worked out by Mandeville. In fact, the Jansenist influence on Mandeville's social thought has been extensively reported .[15] The famous Mandevillian paradox of vices turned into virtues carries the double limitation already made explicit by Jansenists and many French **moralistes** highly esteemed by Mandeville, such as La Rochefoucauld and La Fontaine, that neither are vices noxious nor virtues beneficial. More to the point, vices and virtues cannot be taken at face value. As Horne remarks:

> The general vision stressing the limitations of reason is buttressed by the reduction of apparent virtues to the effects of mere passions.(...) The reduction of virtues to amoral or vicious passions means conversely that characteristics normally thought of as vicious can lead to virtuous actions, at least in terms of public appearance. (Horne,1978:24)

La Rochefoucauld made the point rather crudely:

> self interest speaks all sorts of languages and plays all sorts of roles, even that of disinterestedness. (cited in Horne,1978:24)

In the end, the private sphere surfaces as something quite opaque, for in it reside at least two distinctive classes of driving forces. However, this does not seem to constitute an obstacle for the public to exist or even flourish - provided, Mandeville somewhat prudentially adds, some institutions and more generally laws are made so as to channel our motivations to the public good.

3.2. Compatible-because-conflicting arguments

Despite these complications, some eighteenth century thought appeared to have been more confident about a sort of naturalistic conciliation between the 'fact' (of interest) and the norm of morality. Helvétius,[16] for example, identified self-interest with true virtue and acts of the most enlightened charity. The true virtues which he expected to emerge from interacting self-interested individuals were probably grounded in his commercial vision of social intercourse and its

---

[15] See Horne, op.cit.

[16] Cf. Van Kley, 1987.

underlying view of conflict as diversity. He coined the famous analogy between physical laws and moral laws thanks to the discovery of self-interest and additionally stressed the truly **moral** effects of self-interest.

He was, therefore, at the same time revealing his Jansenist ancestry, since he was working on the Jansenist's rehabilitation of self-love, and sentencing its dissolution, by short-cutting the careful distance Jansenists had traced between self-love and virtue, as Van Kley (1987) remarks.

Montesquieu and James Steuart provided some justification for self-interest in a somewhat optimistic mood, stressing, in turn, suitable **political** effects, according to Hirschman (1977). They claimed, in different ways, the unique belief that an interest-based order was bound to produce positive political results, provided the beneficial spirit of commerce prevails. In particular, Montesquieu believes,

> the spirit of commerce brings with it the spirit of frugality, of economy, of tranquility, of order, and of regularity. In this manner, the riches it creates do not have any bad effect. (Montesquieu, Esprit des Lois, cited in Hirschman, 1977:71)

And, in the Hirschmanian spirit of 'interest countervailing the passions':

> it is fortunate for men to be in a situation in which, though their passions may prompt them to be wicked, they have nevertheless an interest in not being so. (cited in Hirschman,1977:73)

Commercial spirit has the ability to make interests prevail over passions. More to the point, commercial spirit has the faculty, according to Montesquieu, of resisting the harmful tendency of arbitrary power, and defending the inviolability of individuals, through creative devices, like the invention of the bill of exchange and trade **tout court**.

Also, for James Steuart, the intricacy and delicacy of an economic community has remarkable political effects:

> The power of a modern prince, let it be, by the constitution of his kingdom, ever so absolute, immediately becomes limited so soon as he establishes the plan of oeconomy which we are endeavoring to explain. If his authority formerly

> resembled the solidity and force the wedge (which may indifferently be made use of, for splitting of timber, stones and other hard bodies, and which may be thrown aside and taken up again at pleasure), it will at length come to resemble the delicacy of the watch, which is good for no other purpose than to mark the progression of time, and which is immediately destroyed, if put to any other use, or touched with any but the gentlest hand.
>
> A modern oeconomy, therefore, is the most effectual bridle ever was invented against the folly despotism ... (cited in Hirschman,1977:85, quoted from An Inquiry into the Principles of Political Oeconomy)

The interdependence among different people which commerce enables deepens their bonds and enlarges the sphere of common interests, and again restricts the area of political advances of the powers that be.

Here we identify the defense of a commercial community which would arise form the releasing and flourishing of the complementary interests of its members. It was expected, except for some of Montesquieu's doubts concerning the moral character of self-interest, that the sovereign should eventually succumb to the rule of generalized self-interest.

In conclusion, according to the rather stereotyped model of conciliation of interests displayed in 'compatible-because-conflicting' arguments, the many interests are seen as rather directly causing the public interest, as they are understood as diverse, complementary and interdependent interests (and not, say, as antagonistic, rival and redundant ends). The resulting commercial society both requires and yields in its functioning, a suitable moral and political environment.

3.3. Compatible-though-conflicting arguments

Hume, however, imagines another way of connecting self-interest and the public interest, through conventions. He sees conventions as a general sense of common interest (which induces everyone to regulate their conduct by certain rules), which emerges from a (mutually expressed) common sense of (self) interest. He notes that convention

> is only a general sense of common interest; which sense all the members of the society express to one another, and which induces them to regulate their conduct by certain rules. (Hume,1985:490).

It is in one's interest to attend to other's interests, he argues. Yet not in the same sense of the butcher, the brewer, and the baker with regard to the hungry man, which is incidentally, an argument Smith raises against the possibility of benevolence in a large and ongoing society (and, of course, an argument also for a possible justification of selfish interaction).

Hume, instead, deals here with the following conjecture of an individual: 'to leave another in the possession of his goods, **provided** he will act in the same manner with regard to me' One performs an act, that is, in the supposition that the other will reciprocate. The expectation of an action on the part of others conditions one's own action, and conventions render the fulfillment of this expectation a certainty.

> When this common sense of interest is mutually expressed, and is known to both, it produces a suitable resolution and behaviour. And this may properly enough be called a convention or agreement betwixt us, though without the interposition of a promise; since the actions of each of us have a reference to those of the other, and are performed upon the supposition, that something is to be performed on the other part. Two men, who pull the oars of a boat, do it by an agreement or convention, though they have never given promises to each other. (ibidem:490)

We should note that what is at stake here is an increase in the degree of predictability of social intercourse. The resulting order, as it were, is not necessarily harmonious compatibility, or at least it is not predictably so. The idea, rather, is that we have room for a common ground, so to speak. It is in one's interests to pursue common interests, everyone has an interest in the overall regulation, but that is all we can say, the solution being somewhat external to the terms of the problem. That is, we cannot deduce or anticipate it, although we know that whenever it obtains, then certain conditions must have been present.

At this point, it is worth recalling that, although self-interest can have this regulatory role, Hume has some reservations about the widespread use of the category to explain (and justify) social interaction. In fact, Hume warns that the notion of self-interest is not something that we may discover to govern real life interactions, but something that gains meaning in contrast with some disagreeable facts, more basic than it, like factionalism. Interests are only 'less dangerous than a number of more violent and volatile passions', and, he believes, in the words of Holmes (1990), that 'interests, passions and norms conspire together to shape every human action'.

Interest is governed by opinion, so we should not be bewitched by its apparent clarity. It should emerge, as far as possible, from a contrast with non-selfish motivations, social (like benevolence) as well as non-social (like self-destructive passions, and factionalism). The more lively reality of factionalism seems to show how cloudy our minds can be when they fail to grasp our advantage or how blind our acts can be when they fail to express our advantage. In the evocative somewhat Humean words of Holmes (1990):

> Individuals are not clear-eyed about their interests. But they are always impatient of being contradicted. We long mightily and involuntarily to find our own beliefs mirrored in those around us. We grab anyone by the lapels who crosses our path and hammer the unlucky party into agreement because our minds are shocked by contrariness and fortified by consensus. This compulsive intolerance is not a symptom of arrogance but of insecurity. We do not compare alternatives in such situations but clutch desperately at straws. And what we seek is not a strategic ally in the pursuit of power or plenty but symbolic comfort from a fellow believer. This primitive need explains why we are so 'keen' in controversy, something neither self-interest nor even vanity could explain. (Holmes,1990:274)

Hume's recommendation of caution with the notion of self-interest may be said to parallel those of Nicole and Hobbes, advancing a skeptical doubt concerning the presumption of 'virtuous' thoughts and acts. Nicole was drawing attention to the rather moral and external limits of self-interest whilst Hobbes and Hume emphasized the problematic character of self-interest in politics. Among other things, Hume was probably addressing the open optimism of some of his contemporaries with regard to the hailed curative abilities of self-interest. And, by the same token and still paralleling Nicole's attempts, he was implying that self-interest should be seen cautiously, that is, something that we should be attracted to because of the 'disagreeable alternatives', and in this sense, an important piece of an enlightened (critical, for that matter) political argument. Again, Holmes has a nice interpretation:

> Compulsive and impulsive reactions as well as spontaneous sympathies and antipathies provide a foil for 'calculating self-interest', giving sharp contours to that idea. For Hume, and many others, 'interest' is a useless category unless it is reserved for one motive contending with others. He rejects imperialistic attempts to explain **all** behavior by invoking the rational pursuit of personal advantage. Motivational reductionism is unattractive, among other reasons, because it robs 'calculating self-interest' of the specificity it acquires when viewed against a backdrop of selfless urges and thoughtless acts. (Holmes,1990:275)

Again, self-interest arises as something upon which we can manage to improve, as a liberating category. But what Hume probably cannot accept is the existence of a revelation property of it, that might make us see clearly what is going on once uncovered and from this information to lead to the existence of a well-governed community.

So, 'compatible-though-conflicting' arguments stress just the **possibility** of conciliation of the many conflicting interests which ultimately depends on there being some mediate forms of social relations, namely, conventions. These strengthen the overall confidence that everyone shall act out of their interests (instead of more harmful motives) by rendering possible calculation of some sorts.

3.4. The Law that 'hedges us in from bogs and precipices' and Laws of justice

That a sort of morally neutral dimension might emerge form self-interested intercourse has been stressed sufficiently, as suggested by the ideas of self-interest being not intrinsically bad or injecting an element of constancy and predictability in human affairs, or else as a potentially egalitarian claim to the effect that everyone has interests. Still, another quasi-egalitarian consequence has been suggested to the effect that it recommends a generalized distrust in discourses about the common good. In other words, distrust as to the existence of any pre-established notion of harmony mastered by a few - aspects to which we have already referred.

But, then, a morality of 'negative freedom' might also be conceivably associated with it, in that, strictly grounded in self-interest considerations, the legitimate constraints on the self-interest of anyone should be the self-interest of someone else.

Accordingly, we may identify two kinds of laws which somehow stand for these constraints. Locke recommends laws of justice, which are meant not only to give a limitation but also a **direction** to the agent's proper interests. These interests are those which are compatible with the general good. 'Proper interests' are there to give freedom a meaning, a measure. Otherwise, interests might conflate with a 'man's humor', which might, in turn, mean domination or tyranny. Freedom within the law is not, thus, 'to do what he lists' but to 'dispose and order as he lists' **certain** things (the interest derived from a certain property a person has in herself, her actions,

and her possessions, which are all somewhat connected) within the allowance of the law.[17] Following Holmes, the necessity of laws of justice expresses a view that not every 'interest' should be allowed. Let us now turn to this view.

Locke's view of the role of laws of justice is developed to contend with (what is here called) the 'theory of hedges' he ascribes to Hobbes. The Hobbesian 'theory of hedges', according to Locke, makes no claim regarding the content of people's interests, or 'true' interests. It makes rather a disclaimer to the effect that people are allowed to entertain whatever desires or intentions they wish, provided they observe the external restrictions imposed on their acts by civil laws, preventing them from falling off into 'precipices' or being stuck in 'bogs'. Nothing is said, however, regarding the ability of laws to 'elevate' us. Laws of justice, instead, are supposed not only to restrict the range of admissible actions but also to indicate the direction of these actions, enlarging the sphere of freedom which society itself enables. People will then be aware of their true or 'proper' interests, which are those indicated by the 'General good'.

The General Good, hence, seems not merely to overlap with a negative-freedom conception of 'self-interest limiting self-interest' but also to impose from the outside the correct sense of self-interest. Locke's conception of laws of justice, in contrast, implies a constitutive function of the General Good, in terms of the definition of the true interests of people, that is, their interests as citizens. What is the source of the General Good? There seems to exist a place in the social intercourse plan that is independently generated, from the logic of self-interest which is deemed by Locke to provide an unsatisfactory account of the laws as 'hedges'.

Why does Locke require so much from laws of justice? It perhaps indicates that we should think about extra guarantees against the privatization of the public, as somehow suggested in the following quotation by Laslett:

> [law can be positively defined, in Locke,] as the progressive elimination of the arbitrary from political and social regulation.(...) [Locke] develops it into the denial that government can be a personal matter, a matter of will: it must always be an institutional matter, a matter of law. (Laslett,1988:112/3)

---

[17] Cf. Locke, 1988, #57.

To be sure, Smith also stresses the point that the self-interested capitalist would be delighted to become a monopolist (an 'unjust' interest, Holmes adds), thereby revealing his greatest fear of the State's ability to effect this transformation. As a matter of course, this uncovers the related and broader suspicion that the general motivation Smith ascribes to human action, of 'bettering one's conditions', may possibly recommend undesirable (if effective) actions.

Looking into this latter possibility, something else emerges from a reading of the Wealth of Nations. So, again we resort to Smith, in particular the part he plays in the task of liberating self-interest through the role he assigns to the division of labor. This shall constitute an instance of our 'weak compatible-because-conflicting' arguments.

3.5. Weak Compatible-because-conflicting arguments

One might argue that the market itself is expected to constrain non laudable behavior by would-be monopolists. But the remedy may cause illness, Smith warns. In this respect, let us go through the following considerations drawn from Smith's arguments.

Initially let us highlight the skeptical undertones of Smith's defense of selfishness.

In the Wealth of Nations, Smith skillfully explains why we should act (and should suppose others to act) from self-interest. First, because of the extension of the interdependence chains, we cannot feasibly judge the character of the different people with whom we engage in ordinary intercourse. Therefore, it is a prudential maxim not to take for granted the benevolence of these people.[18] Besides, as also indicated by the brewer-butcher-baker case, since many people are involved in these long chains we do not have time to invest in all the special relationships that could ask for their benevolence.

Secondly, in his Theory of Moral Sentiments he also makes reference to the fact that even though we could overcome the consequences of interdependence, our senses could not feasibly jump the

[18] Cf. Holmes, 1990.

barriers between men. We cannot **know** what others suffer, we are each of us best suited to take care of ourselves than of any other person and it is fit and right that this be so.

Self-interest then is something we prudentially assume that others possess, and something that we know about ourselves. But even this knowledge, though positive, carries a negative undertone, our senses **confining** us to take care of ourselves.

Fortunately, the division of labor works in such a 'causally' way that our 'taking care of ourselves' turns on their 'taking care of themselves'. This argument is carefully worked out in the Wealth of Nations, in particular in book I.

Therefore, the division of labor is a way of making the self-interest of some contain the self-interest of others and vice versa, and producing beneficial results. But we should not be blinded by these desirable effects, Smith warns in book V of the Wealth of Nations, as the positive composition effects will generate negative 'decomposition' consequences, so to speak, from its initial inputs. Smith warns about the intellectual, moral or social, and martial disadvantages caused by the division of labor, which is worth quoting at length:

> In the progress of the division of labour, the employment of the far greater part of those who live by labour, that is, of the great body of the people, comes to be confined to a few very simple operations; frequently to one or two. But the understandings of a great part of men are necessarily formed by their ordinary employments. The man whose whole life is spent in performing a few operations, of which the effects too are, perhaps, always the same, or very nearly the same, has no occasion to exert his understanding, or to exercise his invention in finding out expedients for removing difficulties which never occur. He naturally loses, therefore, the habit of such exertion, and generally becomes as stupid and ignorant as it is possible for a human creature to become. The torpor of his mind renders him, not only incapable of relishing or bearing a part in any rational conversation, but of conceiving any generous, noble or tender sentiment, and consequently of forming any just judgement concerning many even of the ordinary duties of private life. Of the great and extensive interests of his country, he is altogether incapable of defending his country in war. The uniformity of his stationary life naturally corrupts the courage of his mind, and makes him regards with abhorrence the irregular, uncertain, and adventurous life of a soldier. It corrupts even the activity of his body, and renders him incapable of exerting his strength with vigour and perseverance, in any other employment than that to which he has been bred. His dexterity at his own particular trade seems, in this manner, to be acquired at the

> expense of his intellectual, social, and martial virtues. But in every improved and civilized society this is the state into which the laboring poor, that is, the great body of the people, must necessarily fall, unless government takes some pains to prevent it." (Smith,1976:781/2)

The ambiguity of Smith's account of the division of labor, as we shift from his positive assessments in book I to the rather negative ones in book V, has deserved a number of comments in the literature. The focus of these comments, to my knowledge, falls on the issue of whether or not there is a contradiction or paradox between the arguments in books I and V.[19]

The more optimistic assessment comes from Rosenberg, and yet it shares with Marx's critical remarks in the Capital the conclusion as to the low 'social and human' prospects of economic progress. Let us briefly go through some aspects of the debate.

Rosenberg rejects the view that books I and V display contradictory ideas regarding the **economic** consequences of the division of labor, at least in what relates to the crucial question of technological progress. According to him, the stupidity of workers that results from increasing specialization is deemed to be perfectly compatible with invention and innovation because, for Smith, the inventive activity is a consequence of at least three distinct components, and to none of these is 'a less developed intelligence' detrimental. In particular, inventive activity is a consequence of: (i) the narrowing of the focus of interest and attention; (ii) the motivation of 'bettering one's condition', and (iii) the complex nature of technological process itself, which renders the inventive activity quite impersonal. As Rosenberg puts it:

> the 'capacity to invent' cannot be assessed or measured in absolute terms; the concept is meaningful only in relation to the complexity of the existing technology and the degree of creative imagination required in order for new 'breakthroughs' to occur. (Rosenberg,1965:133)

Besides, specialization engenders the kind of virtue it needs most, 'intellectual' work or, in Smith's words, the work of philosophers:

---

[19] See Rosenberg (1965), for a reconciliation between the two accounts, and Heilbroner (1973) for the suggestion of a deep and meaningful paradox.

> Major inventions involve the ability to draw upon diverse areas of human knowledge and experience and to combine them in a unique positive fashion to serve some specific purpose. (ibidem: 134)

Rosenberg recalls, for that matter, Marx's critique of Smith's account of the division of labor, in the Capital. To be sure, Marx asserts the correctness of Smith's remarks (originally Ferguson's):

> Being a pupil of Adam Ferguson who showed the disadvantageous effects of division of labor, Adam Smith was perfectly clear on this point. In the introduction to his work, where he *ex professo* praises division of labor, he indicates only in a cursory manner that it is the source of social inequalities. It is not till the book, on the Revenue of the State, that he reproduces Ferguson. (Marx,1977:?)

Still, Marx complains about the situation of the working class and develops further the negative consequences of the division of labor, the cure of which is not to be found, as Smith thought, in homeopathic doses of education sponsored by the state. Marx seems to suggest that it is not only a matter of the 'stupidity' of the working class which is at stake, if we want to go to the real roots of the issue, we should see it as a matter of alienation.

So, it may be said that Rosenberg and Marx's remarks have in common the expectation that the division of labor will work for economic progress, both by increasing the productivity of work and by permitting technological progress. But, now it is obvious that from the point of view of self-interest a justification of economic progress (in non-economic terms) is missing, unless a substantial re-description of self-interest is worked out to the effect that, for example, the alienation of the self is a form of self-accomplishment...!

It is worth quoting Smith from a passage of the Lectures on Justice, Police, Revenue and Arms, where he displays an open disagreement with the happier Montesquieu-Steuart vision on the commercial spirit:

> These are the disadvantages of a commercial spirit. The minds of men are contracted, and rendered incapable of elevation. Education is despised, or at least neglected, and heroic spirit is almost utterly extinguished. (cited in Heilbroner,1973:254)

This image leaves us quite far from our starting point around the almost noble character of interests. Adam Ferguson sees the whole issue as a perverse causality: ignorance is a natural consequence of the division of labor, and at the same time gives birth to industry.

> ignorance is the mother of industry as well as of superstition... [in the industry] the workshop may, without any great effort of imagination, be considered as an engine, the parts of which are men. (Ferguson:182/3)

Ironically enough, the image of self-interest when pushed too far, as required to work out the image of a prosperous and self-governed economic community, gives rise to a debasing engine.

In any case, we have arrived at a powerful source of the 'positive freedom' fears. Self-interest even shaped as true interests, as those sound interests entailed in a modern economic community, may engender unsound interests. It has a disaggregating ingredient of its own, one that requires further and continuous correction.

We are now in a better position to figure out an additional source of Locke's concern about a self-interest based order as well as his uneasiness with the 'theory of the hedges': the order so conceived may generate perverse fruits. Therefore, he concludes that the law of justice should not only be a constraint but also should somehow interfere with or 'direct' our interests. But, then, we would need to lift ourselves by our own bootstraps to reach higher altitudes than our height allows us.

4. A Final Comment

Self-interest passes the test of these two centuries' as an **idea**, a sort of *cogito* of the political thought of modern societies. This seems to mean that it is not a brute fact, a natural or self-evident truth, nor a positive ideal, a firm ground from which normative aims spring, which are the border-lines within which it has been built up. Our ancestors seem to have elaborated on it as a contending piece of political argument. In particular, they made explicit the external limits as well as the internal symptoms of decay connected to this idea.

PART I

SELF-INTEREST AND BEYOND (THE TWO)

INTRODUCTION

In this part, I discuss some images of social order grounded in the behavioral assumption of self-interest where this postulate is justified in virtue of self-interest's ability to produce **efficient** social states or else social states with **equilibrium** properties It is as if the self-interested behavior of the individuals were able to produce a sort of self-transcendence or 'bootstrapping' in the form of coordination of people's plans as well as efficient cooperation. Thus, the title "Self-Interest and Beyond", where it is implied that strict self-interest may carry us beyond ourselves (and as such, surely a normative pretense). I shall be referring here to cells III and IV of matrix 1, the strategic- public and non-public worlds, respectively.

Recall that it was one conclusion in chapter 1 that, from the perspective of its tradition, there is nothing self-evident about self-interest, neither in the sense of its being an obvious reality nor its possessing a universal meaning. Rather, it is to a large extent a construct, or what Thomas Schelling calls a 'vicarious' category, and, in this sense, it does not possess a value of its own, that is, without further reflection on the alternatives. This was taken to provide a major constraint on its use, for both explanatory and normative purposes, that is, even in terms of its ability to explain individual behavior or produce desirable interactional effects. Self-interest is taken to be a residual category (in relation to more basic and less agreeable alternatives) as well as a reflective one (it is a critical standpoint in relation to the received view regarding the meaning of the public interest). As such, the category possesses a limited potential for universality in terms of generating desirable political and moral effects, whose effectuation depends upon the institutional environment. Considered in the light of these ideas, the role of self-interest is largely mitigated, at least as compared to the more optimistic expectations regarding its qualities that have been cultivated by contemporary theories.

In examining a set of these theories, and backed by a view of the political arguments that witnessed the modern birth of the category self-interest, I have come to the conclusion that some contemporary theories misuse this behavioral assumption in that they disregard its obvious limits. This is the case with a branch of the public-goods literature that presents the social order as efficient (and just) cooperation among self-interested or maximizing agents to a

collective action. It is also a misuse of the concept of self-interest when the standard game-theoretical literature proposes a view of coordination that intends to draw exclusively on people's maximizing behavior. Note that the behavioral assumption of self-interest has been translated into the tautological sense of utility maximizing behavior. My examination of both literatures concludes that they fail with regard to their initial plans: cooperation which was to emerge from strictly self-interested behavior seems to need cooperative behavior, and equilibrium in a strategic environment seems to need a non-maximizing explanation concerning the convergence of beliefs of maximizing people. These findings reveal both a moral view and a 'metaphysical' approach to cognitive issues, fundamentally at odds with the value-free starting points of these theories.

In this way, this Part is divided into two chapters. Chapter 2 presents the public-good case of abuse, for it looks into rational choice accounts of the extent to which self-interest may support or provide the public good of political order. An interpretation of Olson's law shall constitute here the yardstick by which the success of such efforts is assessed. Some failures of this approach will emerge, the more interesting being connected with the separation of self-interest and (some ideal of) efficiency.

In chapter 3, the game-theoretical approach is discussed in connection with its potential and the actual tools used to account for social interaction. It is largely a theory of rationality and it is intended to underlie much of public-goods standard theory. In examining game theory's possibilities on the matter, we are indeed pursuing the problems of determinacy of a self-interest based theory, at least in its form as strategic rationality. The separation of self-interest and equilibrium, paralleling that of self-interest and efficiency, is understood as an important indication of relevant epistemic issues thus far inadequately dealt with.

A final word should address the question of why these particular contemporary approaches were chosen in the first place. My answer to that is that they share a conviction and a vision that stand out among alternative approaches, namely, the conviction that we should abstract from behavioral assumptions or postulates concerning **substantive** aspects of human nature, and the vision of social interaction as a complex environment where people's decisions are

affected by each other's decisions and they are aware of this. The importance of 'otherness' in this approach suggests that we are in the world of the 'two'. So, in this sense, examining this framework is tantamount to probing the limits of a deductive instrumentally rational theory to account for what seems to be a more accurate image of social interaction (at least if we contrast it to the image of atoms knocking into each other at random, suggested in the parametric-non-public cell I of matrix 1). Oddly enough, in the course of a solution, this epistemic vision of a world of expectations is exchanged with badly supported behavioral and metaphysical views.

Chapter 2

THE PUBLIC-GOOD LITERATURE'S MISCARRIED 'BOOTSTRAPPING'

1. Introduction

In this chapter I will explore a specific offspring of the notion of self-interest, which has attempted to work it out as the single element from which social interaction stems. In particular, this framework claims that self-interest both accounts for the existence of social order and gives it normative features.

An initial warning must be sounded: I will not dispute here the explanatory merits of this literature but only concentrate on its self-proclaimed normative purposes. In particular, I will dispute the idea, usual in this literature, that social or political order as **cooperation** is the likely result of the rational choice of individuals. Or, in somewhat standard normative terms, that strictly maximizing individuals may find cooperation the reasonable course of action to pursue.

To be sure, early modern political thought has undergone a rational choice re-description. The steps in this process may be described as follows. First, the political order has been understood as a collective action, that is a cooperation of many rational individuals aiming at the achievement of common ends, who are nonetheless free to choose among different courses of actions. Elster (1985) puts the argument into the rational choice form:

> By collective action, I mean the choice by all or most individuals of the course of action that, when chosen by all or most individuals, leads to the collectively best outcome. This course of action I shall ... refer to as cooperative behavior. (Elster,1985:137)

Secondly, contractarian or conventional images of the political order have been translated into collective action **problems**, requiring a rational and determinate solution. Hence, the image of a pre-social state of struggle among competing individuals is portrayed as a dilemma, a choice

between a desirable unanimous (viz. conditional) outcome and an individual best response (viz. unconditional) outcome.

According to Taylor's (1987) definition, a collective action problem exists whenever individual interaction leads to a strictly Pareto-inefficient outcome, or in other words, whenever there is at least one outcome that is preferred by everyone. Elster's (1985) definition further clarifies some aspects involved in a collective action problem, thereby suggesting that more accurate descriptions of the dilemma might be more conducive to its solution. According to him we have a **strong** collective action problem whenever the following two conditions obtain: (i) that 'each individual derives greater benefits under conditions of universal cooperation than he does under conditions of universal noncooperation', and (ii) 'each derives more benefits if he abstains from cooperation, regardless of what others do.' This is, roughly stated, the Prisoner's Dilemma.

Elster goes on to argue that we have, however, a **weak** collective action problem whenever the first condition still holds but the second is split into two further elements, namely, the stability and the accessibility of the cooperative strategy. Therefore, a revised condition (ii)' holds that cooperation is individually unstable and individually inaccessible. The instability of the cooperative strategy choice means that if an individual found himself in a situation of universal cooperation, he would have an incentive to deviate from it. The inaccessibility of cooperation essentially means that an individual has no incentive to take the initiative to cooperate, that is, to move from a situation of universal noncooperation. Elster points out that further interesting collective action problems may contain one or another of the above two elements. In particular Chicken-like situations are typically the case of cooperation-instability, whilst Assurance-like problems are typically the case of cooperation-inaccessibility.[1] As the latter are less 'serious' collective action

[1] The configuration of these games are as follows:

(continued...)

problems, beacuse they lack the element of defection-dominance of the Prisoner's Dilemma, a collective action solution is more promising when it convincingly transforms the situation into one of these characterizations (but not both).

In addition to reducing order to cooperation and changing its origins into a collective action problem, rational choice reading of the contractarian theories typically use the model of the Prisoner's Dilemma as a first approach to represent the situation that might be conducive to political order. They believe the representation in terms of the Prisoner's Dilemma to be faithful to the contractarian description of the state of nature, e.g. Hobbes'. The 'Dilemma' is just the state of nature 'in other words'. Then, they try to show that either this representation is inaccurate and, hence weaker versions of collective action problems would be more cogent, or that it is too static as a depiction of the pre-social ambience. In this latter case, it is suggested that we could use a more dynamic version of the Prisoner's Dilemma. In any case, these contemporary approaches are characteristically a criticism of the contractarian justification for the coercive role of the state, specially that found in Hobbes' Leviathan.[2]

Underlying the re-description of the political order as a particular public good to be provided by some successful collective action is the hypothesis of individual action as strictly self-interested. In other words, this is guided by utility-maximizing considerations alone. This representation has been notorious as the economic theory of collective action, and in particular, since our public good is political order itself, of politics *tout court*. According to this view, order and justice might

---

[1](...continued)

chicken-game

	C	D
C	3, 3	4, 2
D	2, 4	0, 0

assurance-game

	C	D
C	4, 4	3, 1
D	1, 3	2, 2

Prisoner's Dilemma

	C	D
C	3, 3	4, 1
D	1, 4	2, 2

whereas, in the chicken game we have DC>CC>CD>DD, and in the assurance game we have CC>DC>DD>CD, in the Prisoner's Dilemma we have a combination of DC>CC and DD>CD, respectively from the chicken and the assurance games, i.e., of the instability of the former and the inaccessibility of the latter.

[2] See Taylor (1976;1987), Hampton (1989) and Gauthier (1986).

be seen as the outcomes of a non-cooperative game, or, in other words, a situation of interdependence where the agents are aware that the results of their actions are interdependent, and where there is a mixture of conflict and cooperation driving forces.

My claim is that rational-choice descriptions of the problem of the emergence of political order in the sense above cannot find a thorough rational-choice solution. Moreover, it seems to me that this failure indicates both the limits of self-transcendence as far as maximizing behavior is concerned and the fecundity of the so-called external solutions.

My argument is organized as follows. Section 2 looks into an interpretation of Hobbes' account of the origin of political order and Olson's theorem of collective action. The interpretation then suggested is that both accounts stress the idea of order as a predictable environment (rather than a fully efficient cooperation)[3] and accordingly cannot yield a full understanding of the emergence of the political order (or any organization) in 'internal', rational choice terms. This reconstructed Hobbes-Olson formulation will constitute the canonical form of 'external' solutions to the problem of the emergence of political order, which, in turn, is approached 'ecologically', so to speak, rather than mechanically. I take this perspective to be fundamentally faithful to the outlook developed in chapter 1. Section 3 digresses on the 'two laws of the social sciences' that explain the social interaction in terms of the self-interest of individuals - the invisible hand and the 'visible'- to the effect that these are only two different ways of accounting for the same phenomenon of interdependence. In this sense, the two laws ruling the coordination of self-interested individuals indicate the limited reach of the behavioral assumption of self-interest to account for their interaction.

In section 4, the public-goods literature is presented as a number of attempts to approach the collective action target to frame the political order in terms of 'internal' solutions. A repeated-games framework (Taylor's), a Bayesian approach (Tsebelis), and an account that works out the possibility of modification of actors' preferences are then introduced. It is claimed that these internal solutions are incomplete and just add new contexts for the validity of external solutions, namely, comparative-static and dynamic contexts.

---

[3] Following a typology suggested in Elster, 1989.

Section 5 examines solutions to the collective action problem that supposedly, in one way or another, get rid of the assumption of self-interest, what is called here the 'identitarian collective action' solutions; a parallel is set out between this alternative and the problem of factionalism discussed in chapter 1.

Finally, section 6 concludes by remarking on the loss of 'efficiency' in the translation of self-interest into rational choice terms, and the seemingly dangerous prospects of group-identity solutions.

2. Hardin-Hobbes and Olson

It is suggested here that one could gain an insight by considering a different conception of order than that outlined by rational-choice readings of the contractarian undertaking. Moreover, the hypothesis shall be worked out that order thus understood is compatible with Hobbes' vision as well as with an informal version of Olson's argument. Let us turn to the main elements of this idea.

We may think of order either as cooperation or as a somewhat predictable environment, again, in a strong and a weak sense. This double meaning of order comes from Elster's (1989) proposal:

> I shall discuss two senses of social order: that of stable, regular, predictable patterns of behavior and that of cooperative behavior. Correspondingly, there are two concepts of disorder. The first, disorder as lack of predictability, is expressed in **Macbeth**'s vision of life as 'sound and fury, a tale told by an idiot, signifying nothing'. The second, disorder as absence of cooperation, is expressed in Hobbes' vision of life in the state of nature as 'solitary, poor, nasty, brutish, and short' (Elster,1989:1/2)

I will adopt Elster's suggestion of a double meaning of order but depart from him thereafter. In particular, in my judgment, the idea of order as predictability, and symmetrically of disorder as lack of predictability is in accordance with Hobbes image, Macbeth's view being but an exacerbation of the Hobbesian vision. That we need a framework of enforceable rules which somehow enables us to predict others' behavior and, accordingly, yield prospects for cooperation is but a weaker form of the radical unpredictability expressed in the absolute lack of symbolic bonds.

To be sure, the understanding of order as cooperation has the heuristic advantage of rendering it amenable to the rational-choice treatment, in particular by reducing it to Pareto-efficiency or questions of optimality. It is at least disputable, however, whether this is a 'Pareto-superior' explanation, because this requirement is very difficult to match under several collective-action descriptions of the problem of the emergence of order. People may rationally converge to a less-preferred outcome instead of arriving at the better outcome they could feasibly have.

Approaching order as predictability, instead, leaves it open for us to achieve an understanding of the supplementary means whereby predictability is achieved, built, as it were, upon the gaps left by rational-choice explanations. Certainly, we might understand predictability as an equilibrium condition, in the strong sense, that is. Indeed, in a even stronger way, standard economic theory asserts that equilibrium and Pareto-efficiency go together as far as competitive markets are concerned. So, in the provision of private goods under competitive conditions, predictability of individuals' behavior and global efficiency are corollaries, the so-called positive and normative sides of the theory. However, for public goods, these conditions are no longer 'naturally' tied together, their connection being, to say the least, problematic. Besides, it is not unusual to find problems with the justification of equilibrium notions that rely on rather strong informational or rationality assumptions.

Thus, we may understand predictability as the weaker condition of pattern-regularity which might obtain even in the absence of the rationality and information requirements often associated with the equilibrium framework. The view shall be taken that this weaker definition of order is compatible with Hobbes' vision. We may, in this way, find ourselves stressing the mechanisms whereby the predictability of social interaction might be enhanced, and acknowledging that the very instant creation of the enforcement system is somehow external to the rationality of the agents concerned.

Putting it differently, recognizing order and justice as public goods, and trying to solve their production through the solution to a collective action problem is tantamount to undertaking the

task of determining the origin of these 'goods'. But we may conceivably ask, as Hardin (1991) does, is this really Hobbes' or Hume's question? (Is it even worth pursuing?)

In my opinion, the answer is straightforwardly 'no' as far as Hume is concerned. But what about Hobbes? One might take the strong hypothesis that this is arguably not Hobbes' question, as Hardin does. He invokes, as evidence for his interpretation, the logical difficulties commonly associated with Hobbes. In particular, the sort of bootstrapping mechanism, as the very image of the sovereign as materially composed of the bodies of his subjects seems to require, does not look a particularly attractive answer. Why should people abide by covenants, the very need of which is a sign of their structural unfaithfulness? Resorting to sovereign's enforcement is arguably not a genuine way out, for we would still have to explain how basically distrustful people could agree on a rule and a ruler.

So, according to Hardin, the origin of the state is not the focus of Hobbes concern, for it is an insoluble problem at least in rational choice terms. Hobbes' greatest insight, Hardin argues, is to grasp the situation of lawless individuals as a multiplicity of Prisoner's Dilemmas requiring something like an overruling authority for their common or individual purposes to be reached. But the state's origin itself, though correctly depicted as a situation of all or nothing, in the sense that without the state the payoffs would be dire for everyone, and with it an environment of predictability might come about, is left altogether unexplained. Hobbes did not know **how** the order came to be, though he stressed **why** it was needed. It was needed because of the sort of interactions (selfish and even worse) in which lawless individuals would engage. The benefits of it would be the benefits that sprang from a system that provided the means for overcoming the everyday sorts of Prisoner's Dilemmas that people face with normal interaction.

> ...I wish to argue that Hobbes' central problem of political order is not the problem of the Prisoner's Dilemma. We need political order, in Hobbes' view, in part to help us overcome the logic of the Prisoner's Dilemma in our quotidian relations. But the task of creating or maintaining political order is not itself a Prisoner's Dilemma but, rather, a coordination problem. Hence its resolution is not contractarian in the straightforward sense of that term but is conventional. It depends far less (if at all) on what people consent to than on what will work... I also wish to argue that the central problem that Hobbes mastered is not the creation of government from the state of nature but the maintenance of the government. (Hardin,1991:158/159)

This is a quite unusual interpretation of Hobbes' view, for it makes a conventionalist of him, and, if we follow Hardin through, utilitarian as well (a political order is desired because of the prospect of higher welfare levels for everybody[4]). I do not want to claim quite that much, but rather to emphasize what I have already remarked upon in the preceding chapter, namely, that there is an important element of arbitrariness in Hobbes' account of the origin of the civil laws which points to the limits of what is demonstrably true about the constitution of the state, which incidentally, is also an important element in the definition of a convention. He seems to be highlighting the problematic origin of the political body, something very distant from naive images of public interest springing from self-interested action, so that it is hard to conceive the public interest under the clothes of a public good emerging from any choice of individuals at all. A Hobbesian message, it seems to me, is that unconstrained self-interest cannot provide the public good of civil order even if it is in everyone's interest.

Now, in our terminology here, Hobbes' solution to the problem of the origin of the state is said to be 'external', that is, not derived from the rational choice of the individuals living in a lawless state of nature. His idea that any political order is better than none betrays the belief that no choice is possible where there is no rule. So, in this sense, it is trivially true that the question of the origin of the state is external to the individuals' rational choice.

Arguably, Olson's logic of collective action might also undergo a similar reading. There is a certain consensus that Olson's is an economic theory of collective action (Udéhn, 1993, is an important exception), as a theory, that is, of the extent to which self-interest, understood as maximization of one's utility, recommends cooperation or in other words, the maximization of overall utility.

In contrast, I take the view that Olson's is more a 'political' theory of collective action, both in the rather trivial sense that he addresses the question of collective action from an 'external' perspective (in the absence of coercion or selective incentives, large groups will typically fail to provide the

[4] To make this statement sound more than the vague claim that life is better than death, we would need a more sophisticated 'value theory' that would make it possible to measure and compare individual and collective welfare levels. Alas, this is something missing in utilitarian arguments, as Hardin himself openly admits elsewhere (Hardin, 1988).

collective good), and in the less trivial sense that his interesting question concerns freedom and coercion rather than benefits and costs.

Consider that we keep separate benefits and freedom, and costs and coercion, though it might be argued that freedom and coercion could be translated into utilities and accordingly undergo a maximization calculus. But, consider again the question: what is left of the strength of this otherwise powerful argument when one takes into account that maximization is rather the key idea here, for its rule 'prefer more to less' leads people to take less than more of what they might feasibly get? What if we take the trouble to decompose utility into some of its elements and do not take for granted that, say, coercion is straightforwardly deducible from freedom?

So, Olson's question can be posed in a Hobbesian-like fashion as the extent to which free instrumentally rational agents will abide by the restraints they might conceivably impose on their behavior. It implies the rather paradoxical question, which has already been attributed to Hobbes: how much of my freedom should I give away in order to keep some of it? And the answer is 'none', if I could get away with it.

To be sure, Olson's answer varies according to some circumstances of the collective action, the most important being the size of the group. But in so doing he takes the trouble to specify all the complementary mechanisms needed to fulfill the lack of cooperative effectiveness of our rationality. Informally, we may understand him as implying the rather prudential question of how desertion is to be controlled so that the collective endeavor shall not fail altogether, given that where this control is not possible the likelihood of collective action is disastrously low. External support is needed for the collective good to obtain, and additional behavioral assumptions are gathered to color the argument with more vivid and hence cogent images. But the conclusion is strikingly straightforward: we need more than self-interest (even when understood as a harmless utility-maximizing behavior) to sustain cooperation. Incidentally, what I claim here is that it is this view which is implicit in most rational choice arguments for cooperation, notwithstanding the protests to the contrary, as will eventually become clear.

Any internal solution to the collective action problem in its canonic (Olsonian) form shall be considered successful that properly addresses the question of the direct link between self-interest and collective action. It would provide, then, an 'invisible hand explanation' of the emergence of public goods, in particular of political order, and accordingly it should display so-called self-enforcing mechanisms whereby this collective action may arise. Otherwise the solution shall be considered external. Another 'internal' solution shall be examined by the end of this chapter which relaxes either the requirement of **self**-interest (Sen and Pizzorno) or of a thoroughly **rational**-choice (Elster).

My own view is that external solutions have a merit precisely in where they are said to fail. I suggest that we should look at the so-called 'incompleteness' of these solutions as a positive attribute, for it may be a matter of a different way of understanding what is at stake here, for example the extent to which 'social control' of some kind is needed to keep people in line with their 'true interests'. We can think of the external viewpoint as simultaneously suggesting either self-interest's inability to account for the public interest, or the limited reach of self-interest when unconstrained.

3. The two laws of social sciences

Olson has recently remarked that his logic of collective action has actually uncovered a law of the social sciences, indeed, a second law, which follows the 'invisible hand'. This latter is thus stated, in his words: 'sometimes, when each individual considers only his or her interests, a collectively rational outcome emerges automatically'.[5] With regard to the second, Olson goes on to say, sometimes the invisible hand does not hold, 'no matter how intelligently each individual pursues his or her interest, no socially rational outcome can emerge spontaneously(...), only a guiding hand or appropriate institution can bring about outcomes that are collectively efficient'.[6] Sandler (1992) suggests an even weaker form of the law, that 'individual rationality is not sufficient for collective rationality'.

---

[5] Olson, 1992, p.vii.

[6] Olson, 1992, p.vii.

The straightforward cost/benefit calculation, individually undertaken, is not sufficient to sustain cooperation, and this is due to the fact that the benefits of collective action are normally indivisible and non-excludable.[7] The non-excludability condition requires that the benefits of a good be available to all once the good has been provided. The indivisibility condition demands that a unit of the good be consumed without detracting the 'consumption opportunities available for others from that same unit'.[8] These two conditions fix the circumstances where a choice is to take place, in the clear-cut case of collective action in a large group. Individual calculation is then, twofold: on the one hand it indicates the rationality of one's contribution to the provision of the good (if he is any anonymous member of the group), for the obtention of the good is in everyone's interest inside the group. On the other hand, though, it implies second thoughts when the consideration of what this calculus might **contingently** advise enters the stage. In one case, the individual will understand that if he is the only one to contribute, nothing happens, the good is not provided. So, he thinks, 'this conjecture leads me not to contribute'. If, however, others contribute, his contribution is insignificant. All in all, he concludes, non contribution stands out as **the** choice to make.

We feel an underlying conflict between the maximization logic (the best feasible outcome) and the equilibrium logic (the best response outcome) which, to be sure, does not go not unnoticed in recent literature[9] as a conflict between the Pareto-optimum and the Nash-equilibrium concepts,

---

[7] As a matter of technical precision, it is useful to quote Cullity (1995). According to him, "definitions of public goods vary widely, but they usually involve some subset of the following seven features: (i) **jointness in supply**: if a public good is available to one member of the group for which it is public, then it is available to every other member at no cost to that other member; (ii) **nonexcludability**: if anyone is enjoying it, no one else (in the group for which it is public) can be prevented from doing so without excessive cost to the would-be excluders; (iii) **jointness in consumption**: one person's consumption of the good does not diminish the amount available for consumption by anyone else; (iv) **nonrivalness**: one person's enjoyment of the good does not diminish the benefits available to anyone else from its enjoyment; (v) **compulsoriness**: if anyone receives the good, no one else can avoid doing so without excessive cost; (vi) **equality**: if anyone receives the good, everyone receives the same amount; (vii) **indivisibility**: there can be more than one consumer of the good, and each consumes the total output." Olson (1992:p.viii), for instance does not distinguish between (his conceptions of) nonexcludability and indivisibility: "the benefits of collective action are normally indivisible in the sense that, if they are made available to one person in a group they are thereby automatically also supplied to everyone in the group. As is by now widely known, nonpurchasers cannot be excluded from the consumption of the 'collective goods' or 'public goods' that collective action provides." Again, I take the view that a fine characterization of public goods is very important when it comes to explanatory purposes, which is not my aim here. I adopt a characterization with a similar level of generality as that suggested by Olson.

[8] Sandler, 1992, p.6.

[9] Hardin, 1992.

or group-rationality and individual rationality. The way from individual rationality to collective rationality is somehow severed, then.

On reflection, does individual rationality necessarily give rise to collective irrationality in that context? According to Hardin, the problem is rather that we have linked too closely rationality with equilibrium, and this need not to be so. In fact, rationality is quite often at odds with equilibrium:

> That the bonds do not go the other way [from rationality to equilibrium] can be seen from the facts that (1) there can be multiple equilibria with differing payoffs to various players, and (2) there can be equilibria that are disastrous for all concerned and that are Pareto inferior to other possible outcomes. (Hardin,1992:194)

Paralleling Hardin's point, the aspect of interdependence that underlies social interaction may be understood in two distinct ways. Interdependence, in a broad sense, means that the **outcomes** of everyone's actions are linked, and that this arises as an **unintended** outcome of everyone's efforts to achieve their private ends. In a narrower sense, though, interdependence can be characterized as a strategic interaction, in which outcomes and **choices** are related. This deeper sense of interdependence requires that the agents possess an awareness of one another's possibilities, and of the very interdependent aspect of their situation, that are not present in the broader description. Does this distinction recommend two different interpretations of what is in one's interest, or two distinct rationality criteria? Is the rule of self-interest contingent on the broader choice-setting, the one being under the domain of the first law, the other under the domain of the second?

The motivation behind Smith's attributed 'invisible hand' and Olson's law appear to be very different from one another. What we call the 'invisible hand' law was carved very much in the spirit of rescuing 'conflict' from its earlier negative connotations. Indeed it pointed to a potentially cohesive aspect of it, provided it had taken the form of universal self-interest (in the form of disinterestedness in one another's fate). Thus, vices would produce benefits as far as private goods were concerned. However, as soon as public goods enter the stage, the situation changes radically. Vices will no longer produce virtuous results, and what the second law stresses is the potentially disruptive side of conflict. When people have common interests, the achievement of

which turns on their deliberate efforts, the dark side of self-interest appears as a reminder that it cannot be a panacea, since exploitation and opportunism present themselves as forms of rational behavior. Olson, indeed, argues that it may be a matter of domains, the 'first law' applying to the domain of private goods, and the second, to the domain of collective or public goods.

Closer inspection reveals, however, that things may be further complicated if one only considers, that some recent work in economics has stressed the importance of institutions (or hierarchical organizations) to overcome market failures whose origin is precisely located at the micro-level of individual motivation, as in the idea of self-interest with guile and bounded rationality, under particular environments.[10] Moreover, the question of externalities of the market somehow expresses a degree of 'publicness' in the production and consumption of private goods[11] at least in the form of public 'bads', such as environmental pollution. We may also recall here the public bad or the negative externality resulting from the free market in the Smithian vision of an idiot-producing division of labor. Or, the 'even worse' public outcome that would result were private interests to occupy the public space thoroughly, as monopolies would do if the state did not exist.

As a provisional relief to this tension, I suggest a modified version of Olson's predicament in the following form: there is no determinate and straightforward relation between individual rationality and collective rationality. This version has the merit, I believe, of putting 'uncertainty' under focus, without being fundamentally at odds with Olson's law. In support of this view, it is useful to quote Olson's (1971) assessment of his finding:

> The widespread view, common throughout the social sciences, that groups tend to further their interests, is accordingly unjustified, at least when it is based, as it usually is, on the (sometimes implicit) assumption that groups act in their self-interest because individuals do. There is paradoxically the logical possibility that groups composed of either altruistic individuals or irrational individuals may sometimes act in their common interests.(...) Thus the customary view that groups of individuals with common interests tend to further those common interests appears to have little if any merit. (Olson,1971:2)

---

[10] Williamson, 1975.

[11] Lessa, 1993.

Another feature of this latter version is that it mitigates the difference between the 'two laws of the social sciences', in the sense that both fundamentally claim that the relation of the micro-level of individuals' motivation to the macro-level of society or political order is not determinate. In between private- and public- spheres, Smith argues, there exist fortunate compositional effects (which, incidentally are now recognized as the presence of economic institutions[12]), as well as unfortunate ones, what I term 'decomposition effects', that is, the negative externalities of the operation of the division of labor the correction of which is deemed to be the state's affair. Also, Olson adds, there is a fertile terrain for free-riding, in those cases when one can get what one wants without taking the trouble to produce it. Interdependence, the common aspects of both laws, contains uncertainty, for good or evil. The implicit suggestion seems then that we should cast light on the means whereby the bad might be limited or the good hopefully fostered.

Hence, uncertainty or unpredictability, the essential message of both laws, is the crucial trait of interdependence that any problem of collective action should be willing to address. The whole undertaking of the so-called 'internal solutions' may be restated so as to display the self-enforcing mechanisms that self-interested agents are able to produce in order to overcome the basic uncertainty that characterizes their interaction.

4. The public-goods literature: attempts to deviate from the second law

A whole body of literature has considered voluntary cooperation to be a possible outcome of the Hobbesian state of nature. Roughly, it has been suggested that once this latter has undergone a proper description which reveals usually overlooked aspects of human interaction, voluntary cooperation shall follow logically. These aspects that stand out most conspicuously, it is argued, are the dynamic character of this interaction as well as the possibility for people to produce some self-enforcing mechanisms.

---

[12] In a rather broad sense, economic institutions in Smith's economic analysis have been specified by Knight (1992) as implicit patterns of interaction which allow people to profit from specialization gains. The classical reference for a less than full-blown laissez faire-like argument in Smith is Viner (1927). For more recent work, see Song (1995).

So, again, self-interest is expected to make possible the public interest through its ability to repress or neutralize non-social passions. But this is only one side of the story, for it is admitted[13] that in the circumstances of a one-off interaction among self-interested individuals, non-cooperation is still the most plausible outcome, solely on the basis of the existing rationality assumptions. It is then suggested that a case for self-interested, spontaneous cooperation arising without the need of 'the long arm of the state' might be made if we only considered **recurring** interactions of a multitude of actors, which in any case seems to be a more suitable description of ordinary social life than the two-person static scenario.

In the following I shall concentrate on two varieties of solutions that resort to the dynamic component of social interaction, those by Taylor and Tsebelis. Taylor's solution draws on **repetition** while on the whole maintaining the information requirements of classical game theory. The iterative nature of interaction is expected to provide prospects for cooperation because of agents' concerns for their future. Tsebelis's solution, to the contrary, relies on **incomplete information**, stressing the ability of uncertainty among players to enhance the likelihood of cooperation, and therefore deviating the certain and undesirable outcome of overall defection one expects to emerge from Olsonian interaction.

In dealing with Taylor's solution a number of issues will naturally arise. In particular one such issue is provided by Taylor's professed belief that many sorts of institutional solutions that claim to have overcome the collective action problem involved in the emergence of political order have not actually succeeded, for they left unresolved second-order collective action problems.

Therefore, the so-called 'second-order free-rider problem' will deserve a brief mention as well as a brief description of attempts at addressing it as a class of 'less serious' problems of collective action, suc as Heckathorn's.

A more comprehensive treatment is undertaken by the framework of nested games imagined by Tsebelis. It will nonetheless lend evidence for my hypothesis that increasing attempts at internal solutions of increasing degrees of collective action problems end up by displaying the 'irrationality'

---

[13] Taylor (1976) and (1987).

of obsessive rational choice, for it blurs any possible dividing line between what it is and is not rational to do in any particular case.

My conclusion will be that the Hobbes-Olson problem is not overcome, except when an extra-rational move is also implicitly assumed, such as additional behavioral assumptions to self-interest (that some people are conditional cooperators, or cooperators, in Taylor and Gauthier) and implausible forms of 'hyperrationalism' (as in Tsebelis). These authors implicitly enlarge the area of validity of the Hobbes-Olson predicament, by specifying the need for extra-rational support for cooperation in non-strictly static contexts, and in this sense add contingent reasons to its legality (Lessa, 1995). A public good with the characteristics of the political order shall be provided where it relies less uncompromisingly on bare self-interest or individual rationality.

4.1. Taylor

Taylor (1987) provides a well-specified definition of an 'internal' solution. It goes as follows:

> **Internal** solutions neither involve nor presuppose changes in the 'game', that is, in the **possibilities** open to the individuals (which are in part determined by the 'transformation function', specifying how much of the public good can be produced with a given contribution), the individuals' preferences (or more generally **attitudes**), and their **beliefs** (including expectations). (Taylor,1987:22)

And, subsequently he provides a justification for adopting an internal solution:

> I shall take the view that the internal solution is the **basic** one, in two connected senses. It is, first, the only one which is complete in itself. All the external solutions presuppose the prior and/or concurrent solution of other problems, usually (always?) of collective action problems. Many of them, for example, involve the use of threats and offers of sanctions, and the creation and maintenance of the sanction system entail the prior or concurrent solution of collective action problems.(...) The internal solution is basic in a second sense: until we know whether a solution of this kind is possible and what form it will take, we cannot say what work, if any remains to be done by other putative solutions. (ibidem:22)

The criticism in the first part of the statement above is meant to apply to many solutions such as 'political entrepreneurs' (Hampton), 'property rights' (Buchanan), and 'norms' (Ullmann-Margalit,

Elster). That is to say, they all require a second-order self-interest based derivation of these mechanisms, according to Taylor.

So, admittedly, a failure in his solution to the iterated N-person Prisoner's Dilemma, which he deems to be an accurate description of the Hobbesian state of nature, will display the difficulty in solving the collective action problem altogether. For the recourse to 'other putative solutions' still carries the burden of leaving unsolved prior collective action problems. Moreover, one such failure weakens the criticism of the coercive role of the state, which is the major normative purpose of Taylor's efforts:

> The most persuasive justification of the state is founded on the argument that, without it, people would not successfully cooperate in realizing their common interests and in particular would not provide themselves with certain public goods: goods, that is to say, which any member of the public may benefit from, whether or not he or she contributes in any way to their provision.(...) The Possibility of Cooperation is a critique of this justification of the state, and the heart of the critique(...) is a detailed study of cooperation in the absence of the state and of other kinds of coercion. (...) Hobbes' Leviathan was the first full expression of this way of justifying the state. The public goods he was principally concerned were social order - domestic peace - and defense against foreign aggression. Without these, very little else that was worth having could be had. (ibidem:1)

Indeed, Hobbes' account of the origin (and justification of the existence) of the state relies, according to Taylor, on his description of the state of nature as a one-shot Prisoner's Dilemma game. Since this game does not offer prospects for cooperation, although it also displays players' willingness to coordinate on a beneficial cooperative outcome, Hobbes' conclusion is that a coercive mechanism, the state, is welcome which might enforce the contribution of the individuals.

What, Taylor asks, is the possibility of cooperation among a large number of individuals in the absence of coercion? It turns, in the first place, upon the possibility for the cooperative strategy to be the likely action of these individuals or of a sufficient number of them. And this possibility can only be defended if the strategies are seen as somehow interdependent, in the sense that an actor's strategic choice is contingent on the others' strategy choices.

To be sure, this is a condition that is also present in Olson's setting of collective action as an actor's choice of strategy is contingent on what he thinks others will do as when he considers that being in a large group his defection (or his contribution) will go unnoticed. And since each individual will have the access to the public good anyway, Olson concludes that only a coercive mechanism or some separate incentives other than the 'common interest' itself will motivate individuals towards cooperation. Yet, Taylor's idea of a contingent strategy harbors a somewhat different element, namely, that interaction involves a **sequence** of plays among several players. Let us put together the main parts of his argument, one by one.

As is well known, Hardin (1971) analyzes Mancur Olson's collective action problem as a N-person Prisoner's Dilemma. This has entailed a slight modification of Olson's logic, for the non-cooperative outcome appears as a result of the operation of the dominance logic: regardless of what others intend to do, noncooperation stands out as the individual optimizing strategy. This manoeuvre has had the effect of depressing still further the already low prospects of cooperation contained in Olson's idea of individual invisibility in a large group, for ever since his law has implied that individual rationality leads to collective irrationality. I conjecture that the Olsonian emphasis on invisibility (or the marginal contribution) instead of dominance arguments was conducive to his solution to the collective action problem. Collective actions are more likely to succeed where visibility is quite high (as in small groups), but they need extra restraints in so far as large and intermediate groups are concerned.

In any case, Taylor builds upon the work of Hardin, in the sense that he proposes a solution to a version of the Prisoner's Dilemma (PD) supergame, namely the iterated N-person PD. Though Taylor in an earlier work (1976) had already modeled the iterated PD, the most famous results are attributed to Axelrod's (1984) computer tournament. It consisted of an indefinitely repeated number of two-person PD games, all participants (a number of philosophers and social scientists) playing against one another; the winning strategy was 'tit-for-tat', that is, cooperating in the first round and then imitating what the other player has chosen. Axelrod's conclusion is that

cooperation is possible whenever two conditions obtain: 'that the cooperation be based on reciprocity and that the shadow of the future be important enough.'[14]

To be sure, Taylor argues that many collective action problems take the rather weak form of the chicken or assurance games, which are games with multiple equilibria, requiring some sort of coordination to select a specific equilibrium. But were one to make a game-theoretical problem out of the Hobbesian state of nature, and, accordingly, to search for a determinate and unique solution to it, one, in particular, that reconciles individual and collective rationality, the task is surely to solve the Prisoner's Dilemma, as Taylor goes on to argue. So, an internal solution requires that the problem itself be modeled under the Prisoner's Dilemma conditions.

What was then missing in the Hobbesian description of the state of nature that obstructed the way to a solution? According to Taylor, Hobbes' state of nature was inadequately static, and being so, it unnecessarily narrowed the sense of self-interest to what Taylor calls 'eminence' The motivation of eminence engenders a game of difference where payoffs for one player increase with the difference between his and others' utilities, instead of his welfare or utility level alone. Had Hobbes considered the dynamic nature of interaction he would have concluded that a less harsh sense of self-interest would come about which could somehow neutralize eminence. For this to occur, Taylor goes on to argue, it suffices that we introduce a condition that. incidentally, enhances the realism of our model, namely, repetition or 'time'.

Interactions among individuals motivated by self-interest might engender voluntary cooperation if only a number of people give positive value to future payoffs. PD's logic, let alone Olson's, might be overcome through noncoercive means, and hopefully rationality might be unleashed and the dilemma solved if only time were considered.

Certainly, as Axelrod had already pointed out, reciprocity is also an indispensable part of voluntary cooperation. So, rephrasing Taylor's statement, reciprocity (or the use of contingent strategies, in his words) is more likely where repetition occurs.

---

[14] Axelrod, 1984, p.173.

So, time is a necessary condition for cooperation but not perhaps a sufficient one. That is, if people give some value to future outcomes they may come to consider the possibility of cooperating. But this possibility is highly contingent on similar plans by a number of other people. Given this uncertainty, overall defection is always an equilibrium, as Taylor has shown. But still, we are interested in other possible nondefective equilibria, that is, at least on their logical possibility.

This possibility, as becomes clear in Taylor's analysis, is contingent on the interplay between two kinds of players, namely, unconditional cooperators and conditional cooperators. Further restrictive analytical conditions stipulate, the desirable limit to time discounting, the proportion of conditional and unconditional cooperators, the number of repetitions (that should be indefinite, otherwise defection dominates, as he points out), and the number N of the players, which being too large would not impose costs to free riding.

So, summing up the requirements of a solution, we need two kinds of cooperative behavior for cooperation to emerge from a population, or, to put it differently, we need some players that define their strategies conditionally. Additionally the proportion of cooperators by nature and by circumstance must be fixed in equilibrium, and time discounting should be smaller the larger the number of unconditional defectors. Last but not least, we should have a 'not too large' population to enable fake cooperators to be detected.

The usual comments on Taylor's analysis apply here, his own comments coming first. He concludes that PD is not overcome with the introduction of time, although multiple equilibria seem to be more hopeful an outcome than a unique equilibrium where all players defect. Voluntary cooperation is more likely to emerge in small groups than in larger ones, where the usual Olsonian caveats apply.

In other words, cooperation is more likely to emerge from models where substantial modification of the behavioral assumptions is introduced, cooperation being postulated instead of deduced as would be required by a wholly internal solution. As for the introduction of the repeated game framework as a way out of the collective action problem, we had better quote Aumann who notes

that repetition 'fails to eliminate the ultimate inefficient outcomes - the payoffs to a strategy of mutual defection.'[15] But Hechter (1992) also complains against the over optimistic informational assumptions of the repeated game framework, for it is blindly confident about complete and perfect information. Opposing the assumption of complete information, Hechter argues:

> To impute strategies and payoffs to other players, individuals would need to know (1) each other's action resources and constraints (in order to arrive at a mutual understanding of their respective set of available strategies), (2) the causal consequences of all n-tuples of strategies (in order to know the set of attainable outcomes), and finally, (3) each other's perceptions and subjective valuations of these outcomes (in order to construct the payoff matrices that would then be subjected to the mathematical analysis of game-theoretic solutions). (Hechter,1992:36)

Hechter goes on to point out that the assumption of perfect knowledge implies that players have zero monitoring costs which is too unrealistic, especially in the n-person game.

The startling thing about Taylor's model, it seems to me, is that even having to justify such strong assumptions as its behavioral postulates and the informational framework, it still ends up with a rather weak solution, in the form of multiple equilibria. Of course, with multiple equilibria collective irrationality may still be individually rational, for equilibrium reasons: every individual has a good reason to take an equilibrium strategy that, nonetheless, may match in a non-equilibrium strategy profile.

A final comment concerns the introduction of time in Taylor's argument. As time is incorporated as a time-discounting rate we are perhaps justified in thinking that Olson's static approach was not completely left behind. Taylor himself acknowledges that his treatment is not properly dynamic but comparative-static. Yet, it may be worth looking into this, I shall argue.

The possibility of cooperation arises hopefully from interaction over time. This latter, in turn, is translated into a discounting-rate. However we are not sure as to how this discounting-rate is itself determined, in other words why people come to assess the future in an optimistic mood. An idea that offers itself to scrutiny is that future outcomes are worth having because we hope to

---

[15] Cited in Hechter, 1992, p.36.

have built a bridge with others towards that future, in other words, our positive assessment as far as future is concerned also depends on the already existing degree of cooperation. In this sense a low time-discounting instead of **explaining** cooperation may be **assuming** it. That time is somehow exogenous and objective (replacing Leviathan?) is confirmed by Taylor's postulation of an invariable discounting-rate.

In short, Taylor's efforts seem not to overcome the Olsonian prediction as to the higher probability of cooperation in a small group than in a larger one, due to the additional precautions he deems necessary in order to make sure that conditional cooperators will actually cooperate, even in the presence of the favorable circumstances for cooperation, namely, some preference for future outcomes and a variety of types of actors. These precautions refer mainly to informational conditions, or transparency, less likely in large communities, more abundant in smaller ones. In the end, this is the possibility of cooperation (or of 'anarchy'). Cooperation exists where there already exists a genuine willingness to cooperate or mechanisms to detect non-laudable behavior.

Hence, as a final word we cannot help concluding so far that the introduction of realism into Olson's model in the form of repeated interaction cannot avoid achieving the same results as his own, leaving an internal rational-choice solution to the collective action problem of creating the political order in the same irreducible plane of irreality as in its canonical, Olsonian form. Indeed Taylor's efforts seem to have added second-order reasons to the Olsonian logic: in the absence of coercion or separate incentives, neither will non-strictly static contexts produce cooperation, except if small groups are assumed.

As a side point, Taylor's multiple equilibria and informational burdens might well open a case for the banality of a collective action **deductive theory**. This is the route taken by Hechter and Elster. They claim for more dense descriptions of a particular collective actions dilemma and for the grasping of relevant norms in each case. Hechter (1992), for example, points out the impossibility of the construction of a payoff matrix if we stick to realism and drop the traditional informational assumptions. Elster (1985), in turn, suggests that we should work out rather weaker conceptions of collective action problems than the Prisoner's Dilemma (Chicken and Assurance like descriptions of it), which are more realistic and, accordingly, give room for contingent actions and

prescriptions for actions. Also Udéhn (1993) suggests an approach for collective action that encompasses mixed motivations and the heterogeneity of people. He argues that in the extended literature stimulated by Olson's pioneering Logic, much sight has been lost as to the sociological rather than analytical aspects of the problem, which were present in Olson's original formulation.

But this is not the view I shall pursue, being still interested in looking further into the justificatory role of self-interest in political arguments, in particular in those concerned with the constitution of the political order as resulting from self-interested undertakings.

4.2. Second-order collective action problems?

The second-order free rider problem, pointed to by internal solutions as a weakness of external solutions, is not a fatal critique, or so Heckathorn and De Jasay seem to claim. De Jasay (1989) rejects the free-rider **problem** altogether whilst Heckathorn (1989) sets out to demonstrate that the two levels of free-riding are distinct to the effect that cooperative behavior is more likely at the second level.

In the first place, what is the second-order free riding problem? According to Taylor's suggestion, it is the problem of building a system of incentives (negative or positive), stemming entirely from the self-interest of participants to it, in order to solve the problem of provision of the first level of public goods that every participant is directly interested in. It is nearly equivalent to Hardin's vision of participants to a collective endeavor being able to pull themselves by their own bootstraps in order to change the incentives they face. It seems logically impossible. Recalling Hardin's words, such an incentive system is welcome, to be sure; moreover, it is all the more logical; nevertheless, it is thoroughly impractical.

However it has been argued that second-order collective action problems are conceivably 'less serious' than first-order ones. Heckathorn (1989), for example, states that we all have two distinct sets of motivations, namely, what he calls 'inclinations' (or direct concern with one's welfare) and 'regulatory interests' (concern with others' behavior). Our inclinations are normally in conflict with our regulatory interests in the sense that the former tempt us to free-ride whilst the latter advise us to engage in the construction of a regulatory system that rewards cooperation and punishes

defection. He concludes that rational people tend to cooperate in the second order and to defect in the first order collective action problem, leading to a behavior of 'hypocritical cooperation'.

Thus, Heckathorn frames the game played in two levels by a person considering the four feasible alternatives she is to face when deciding whether or not to engage in the provision of a regulatory scheme, and whether or not to contribute to the first-order cooperative enterprise this regulatory framework is supposed to constrain. He will conclude that a likely dominant strategy of hers is to engage in what he calls 'hypocritical cooperation', which means to cooperate in the construction of the regulatory or coercive scheme, and to defect in the first order choice. That is, to make for the cooperation of the others through an enforcement framework, but to deviate when it comes to pay the costs of the immediate public good. He then takes the route of exploring the dynamical benefits of this strategy insofar as it is expected, as time goes by, to enhance the likelihood of 'full cooperation' (cooperation in the two levels) itself.

Without going into the details of his model, I will just note a basic failure in his argument that the second-order interaction is governed by what he calls regulatory interests, which in my view is misunderstood as an interest in the behavior of **others**. It is odd that he does not see that the actor himself is someone who will be subject to the second-order rules as well, and so defection, on this account, should be his more likely strategy.

For that matter, let us briefly refer to Elster's Ulysses as a correct understanding, in my view, of the kind of problem here at stake. He says that we need to bind ourselves, but how? In the myth it is easier to grasp how a solution may be brought about through one's acting on one's external environment. So Ulysses asks

> ...you must bind me hard and fast, so that I cannot stir from the spot where you will stand me ... and if I beg you to release me, you must tighten and add to my bonds. (cited in Elster,1979:36, quoted from the Odyssey)

In our daily interactions insofar as our welfare is concerned, namely refraining from one's addiction and similar cases, it is also arguable that binding oneself is the best strategy, though it

does not cease to be problematic.[16] But these are clearly not cases of **public** goods, and when it comes to the yield of really public goods, as refraining from littering and the like, Elster (1985) argues for norms, which are not deducible from rational choices. According to him, the strength of 'rational choice' reasons for abiding by social norms (in our case, norms of cooperation) fades away as soon as increasing layers of 'choices' are supposed, the strategic sense of our following social norms becoming less clear.[17] In particular, to follow a social norm for strategic reasons seems to be self-defeating:

> Unless rules were considered important and were taken seriously and followed, it would make sense to manipulate them for personal benefit. If many people did not believe that rules were legitimate and compelling, how could anyone use these rules for personal advantage? (cited in Elster 1989:128, quoted from Edgerton)

Ulysses's problem, when it comes to social cooperation in rational choice terms, turns out to be whether 'rational, selfish, outcome-oriented persons will be able to cooperate', and this will happen 'only if each of them is able to impose the cooperative strategy on **his** successive selves.'[18]

De Jasay's (1989) argument in turn leads to a prescription of tolerance with a 'spontaneous' proportion of 'free riders' and 'suckers' brought about by free-exchange. This prescription comes with a threat: let this proportion be freely achieved otherwise (that is, with social-democratic intervention) the proportion of free-riding will increase with a vengeance (that is, with the expansion of the public goods domain entailed in a social decision rule and the consequent perverse incentives to free-riding). Voluntary provision of public goods can be consistent with narrow self-interest only if we accept a spontaneous rate of exploitation, warns De Jasay. In the trade-off between 'liberal' frame (free-exchange) and democracy, De Jasay sides with the former.

On reflection, we may find an argument highly unsatisfactory whose justification relies on the scarcely defensible claim that what is 'natural' ought to be set free. Our suspicions concern the

---

[16] Elster, 1979, p.37.

[17] Also Elster, 1989.

[18] Elster, 1985, p.147, my emphasis.

supposed naturalness, let alone freedom, of so-called 'free-exchange', since market sanctions fall heavily on the shoulders of not-always-conformed suckers.

On De Jasay's account, natural exploitation may lead to efficiency, which incidentally yields a curious description of the so-called 'free exchange', and on Heckathorn's account hypocrisy is expected to produce social virtues. These claims are much stronger than the arguments for self-interest I want to examine here, and for this reason I shall not dwell on them.

4.3. Nested games

In the branch of solutions to collective action problems that take into account new incentives that evolving interaction may provide, the endogenous approach to institutions seems to occupy an important place. In addition to the notion that institutions may be produced with the purpose of changing the situation the actors are involved in, i.e. as self-enforcing mechanisms (like the already mentioned entities of 'political entrepreneurs', the system of property rights, or norms in the sense of Ullmann Margalit[19]), and in an attempt to escape the hardship involved in such theoretical efforts, Tsebelis worked out an original way of understanding collective action, which incidentally is highly normative in character.

This research program calls for a dynamic approach to social interaction, contrastingly with the usual static and comparative-static ones, relying heavily on a strong version of rational choice. It also claims to criticize the informational setting often required in both static and repeated-games framework, by assuming incomplete and imperfect information instead and, thereby, discovering an additional incentive to cooperation, namely that provided by uncertainty itself. Its starting point, however, is the assumption of full-blown rationality on the part of the actors (that is, in this context, Bayesian rationality), which is actually quite a strong rationality assumption. How do these two elements relate, namely uncertainty and full rationality?

Tsebelis seems to suggest that we should separate uncertainty from indeterminacy, by asserting that actors possess adequate resources to deal with the uncertainty that normally conditions their interactions. However, he contends, theoretical undertakings very often misleadingly identify

---

[19] Norms as approached by Ullmann Margalit derive from strategic reasoning.

suboptimal behavior in real life actions that are actually **optimal**, thus underestimating actors' rational capacities. External observers, like the theoreticians, he goes on to argue, often fail to perceive the players' **real** situation, in the sense that they overlook the fact that players' choice of strategy in one particular arena is quite frequently contingent on what happens in other arenas where they are also participating. That is, relevant contexts to players' decisions are unjustifiably omitted, and with this omission one misperceives their **real** payoffs. Another aspect of the decision setting that is often underrated is the fact that players may also be involved in a sort of institutional choice, where the very institutional design, that is the rules of the game, is being decided, and decided simultaneously with the decisions in an immediate arena. This latter aspect casts light on the element of variability of the very constraints to be imposed on an optimizing choice of strategy.

Taken together, these contextual and institutional aspects modify the direct game the player is concerned with and may lend prospects for cooperation even in the traditional Prisoner's Dilemma game. Why so? According to Tsebelis it is due to the fact that this framework, what he calls the nested games framework, relaxes the traditional informational requirement of classical game theory, of complete and perfect information. If the game that players are involved in is not taken in isolation, contextual and institutional-design factors being needed for their full intelligibility, Tsebelis sets out to demonstrate that these factors while introducing some uncertainty as to the likely choices in the primary arena also make for the possibility of cooperation among players. Under the framework of the Prisoner's Dilemma, with the usual informational requirements, the rational outcome is mutual defection. However, if we simply introduce less than complete information even in the single-shot version of this game and additionally admit that players might communicate, an entirely new range of strategies, contingent or correlated, may appear. In particular, the consideration of a larger context than the immediate arena or the single game may render other strategies in a given set more relevant. Moreover, the consideration of a modification in the rules of the game may give rise to an entirely new set of strategies.

Correlated strategies mean that each player chooses his strategy according to the opponent's strategy. Once communication is permitted and players can coordinate their strategy choices, 'the likelihood of different strategies varies with the **magnitude** of each player's choices' and not only

with the usual preference ordering. In particular, the likelihood of cooperation increases with increasing payoffs for cooperation and decreasing payoffs for defection. The important component here is players' expectations concerning the use of cooperative (or otherwise) strategies by their competitors, without a change in the basic preference structure being required.

Since players are now interested in maximizing their expected utility their choice of strategy will also depend on the probability they assign to the various courses of action which may be undertaken by others, taking into account the pre-play talk. In particular, the likelihood of cooperation is said to increase with the probabilities **p**, of instruction (that a cooperative strategy shall be followed by others), and **q**, of retaliation (that a defection is likely to be punished).

Of course, it all depends on the binding force of the pre-commitments. Promises (of cooperation if the other cooperates) or threats (of punishment were the other to cheat on the pre-play agreement) are more understandable and trustworthy the more developed are communication as well as coercive institutions, and, obviously make sense only in the context of repeated interactions.

Thus, the case for contingent strategies is strengthened by the consideration by Tsebelis of iterated games. He undertakes to show that within the setting of repeated games, contingent strategies are highly likely to emerge. In the context of repetition, Tsebelis argues, players' interests are furthered to the extent that they maximize their utility over the entire period of their interaction. However, according to him, the usual repeated-games framework misses the point in stressing the **number of repetitions** as the crucial variable. For what matters is not whether repetitions are finite or infinite, because in both cases 'always defect' can be shown to be an optimal strategy choice, as Fudenberg and Maskin's (1986) formalization of the 'folk theorem' with respect to iterated games proved. This theorem states that 'any individually rational outcome can arise as a Nash equilibrium in infinitely repeated games with sufficiently little discounting'. The point is that Fudenberg and Maskin have also shown the theorem to be valid also for a finite (provided it is sufficient large) number of repetitions, if there is incomplete information.[20]

---

[20] Incomplete information is sufficient to eliminate the distinction between a finite and infinite number of rounds. Fudenberg and Maskin proved that cooperative strategy could be the equilibrium outcome of an iterated game when

(continued...)

What matters here is the fact that the likelihood of cooperation is enhanced by the players entertaining some reasonable hope (backed, as it were, by both credible promises and threats) that (some) others will cooperate conditionally on their cooperation. So, contingent strategies are worked out in Tsebelis's approach as the enabling devices that create prospects for cooperation.

The deductive framework is developed in the first chapters of Tsebelis's book, and it aims to work out the logical conditions for cooperation to emerge. However, this deductive framework is not only to perform the role of setting out the **ideal** conditions for rational choice, for Tsebelis also makes the **empirical** claim that real actors are always maximizing and optimizing agents. In this sense, rational-choice framework is not only supposed to **establish** the ideal conditions for individual rational choices but also to **uncover** the optimality of apparently non-optimal behavior by real actors. This latter claim is heavily dependent on the successfulness of his approach to institutions, for contingent strategies are more effective where coercive and communicative institutions obtain, that is where promises and threats are both known and credible. Let us turn to this now.

Tsebelis argues, then, that players' payoffs are also variable because the games they play are sometimes nested in a larger game whose purpose is precisely to change the rules of the stage games. This is a game that intends to change the rules of other games, or institutional design. Actors are very often simultaneously involved in these games, and their aim, in so doing, is to enhance their outcomes, either by actively participating in the shaping of **efficiency** institutions (those that offer prospects of Pareto-improvement), or by engaging in **redistributive** institutional design (that offers prospects of improvement of the position of some groups to the detriment of that of others). Putting stress on efficiency institutions, Tsebelis argues that these are institutions that normally help to push outcomes towards the Pareto frontier, solving general problems of coordination which are highly likely to appear given our bounded rationality (limited capacity to foresee the consequences of one's actions as well as others' actions in general), opportunism ('discrepancy between ex-ante promises and ex-post behavior'), and asset specificity (the assets

---

[20](...continued)
the number of iterations is either finite (if there is incomplete information) or infinite. (Tsebelis,1990:76)

are somehow tied to certain actors,[21] so that actors have continuing interests in the identity of one another).

So, we are told, we need institutions because we have the kind of problems we cannot automatically surmount. Yet, Tsebelis makes it quite clear that his approach to institutions intends to address the production of institutions as a problem of maximization under constraints: 'actors maximize their goals either by changing their strategies or by changing the institutional setting that transforms their strategies into outcomes.'[22] It is not, then, a matter of 'emergence' which is here at stake, but rather a question of conscious and intentional maximizing behavior aiming at producing the maximizing institutions. Tsebelis, then, compares institutions to investments, a personal engagement in a long term enterprise whose fruits are to be collected later in time.

Now, uncertainty as to the consequences should also apply here in the context of the construction of institutions, just as it applied in that of multiple arenas. However, oddly enough the fact that institutional design is a likely choice under uncertainty receives a very limited emphasis. Unintended consequences are deemed to be a side issue, for what is more important according to Tsebelis is the fact that institutions are created and modified following conscious and intentional actions, that is rational choices on the part of political actors. However, the rational choice of the rules-game, we may object, cannot itself be nested in a bigger game, that is, cannot benefit from the heuristic advantages of incomplete information, as with the primary games, unless we enter into an infinite regress. Additionally, how do we reveal the rationality of the institutional choice without recurring to some functionalist-like argument to the effect that a specific institutional design is said to be rationally chosen because it has proved to be beneficial to our coordination? Tsebelis's purpose, however, was to show institutional design as the **result** of conscious actions by rational actors aiming to overcome their coordination failures.

Tsebelis certainly acknowledges that there is a discontinuity in the treatment between the two parts of his book, his account of institutions being far more inductive than the 'games in multiple

---

[21] This definition was suggested to me by Andy Lewis.

[22] Tsebelis, 1990, p.96.

arenas' framework. Yet we should try to assess the extent to which this failure affects the endeavors of an endogenous solution to institutional design and change.

My supposition is that it keeps a striking resemblance to Hobbes-Hardin and Olson's difficulties. Again, coercive (or other regulatory) mechanisms (or institutions) are admittedly needed to overcome likely failures of coordination among rational-choosers. Institutions are generally considered rational, Pareto-improving for that matter. It would therefore be wise for people to engage in the production of these institutions. But, still, why should we expect these people to do so? How is it possible that limitedly-rational, opportunist people overcome their faults in order to build the very mechanisms that would prevent them from failing to be rational? If they could do it they would not need it.

A final related doubt refers to the general possibility of building a thoroughly comprehensive rationalist approach. In my view this is not possible (after Popper's (1945) statement).[23] Hyperrationalism is but a childish disease of rationalism.

Keeping in line with our former trajectory from Hobbes-Hardin, through Olson, to Taylor we may conclude that further contingent reasons have been added to Taylor's argument by Tsebelis. It is not only that under complete information, (single-played or iterated PD-like), cooperation between rational, self-interested, independent players cannot develop,[24] but also that under incomplete information and stronger rational-choice requirements, cooperation is only achievable with the aid of coercive and communicative institutions whose creation and modification we have thus far been unable rationally to deduce.

4.4. Modification of actors' preferences

This section has the purpose of briefly mentioning alternative attempts to deviate from Olson's law, that postulate the possibility for people to so transform their preferences over actions as to favor cooperation.

---

[23] For a detailed discussion of the Popperian argument, see Part III, section 8.

[24] Tsebelis, 1990, p.102.

One such account assumes the possibility that people rationally choose to transform themselves into cooperators, thus moving from their narrower interests. Preference changes are not assumed but derived, as for example, a consequence of individual adaptation to situational incentives, as has been suggested by Raub and Voss.

Hence, preferences may themselves be a matter of choice, of rational choice. Morality has also been assumed as a matter of rational choice by Gauthier (1986). The willingness to cooperate is itself assumed to be a matter of choice, motivated by individual interests. Rationality is then taken to be twofold: it is both related to self-interest and transcendent to it. Morality, accordingly, is to be understood as a rational (utility maximizing) constraint on the pursuit of self-interest, and 'rational constraints on the pursuit of interest have themselves a foundation in the interest they constrain'.[25] We note here a similarity with the sort of bootstrapping involved in the related arguments for long term self-interest.

Morality emerges, according to Gauthier, as a chosen constraint. 'Morality [impartiality]... can be generated as a rational constraint from the non-moral premisses of rational choice.'[26] His conception of rationality, though impartial, he claims, is not universalistic but maximizing; it does not, hence, require any 'veil of ignorance' assumption. Gauthier claims that under certain structures of interaction (like the state of nature) that may be modelled as a bargaining problem agreed mutual constraint emerges as a rational response.

An important question which his approach needs to address is how the *ex-ante* hypothetical agreement on mutual constraints may turn out to be an *ex-post facto* one. In order to achieve this purpose, Gauthier defines four core concepts: the concept of market as a morally free zone; the concept of minimax relative benefit, that is supposed to solve the bargaining problem; the concept of constrained maximization, or conditional cooperation; and a proviso, which is a condition regulating the initial bargaining position. These core concepts are the tools with which he works out the necessary and sufficient conditions for cooperation to emerge among self-interested people.

---

[25] Gauthier, 1986, p.2.

[26] Gauthier, 1986, p.4.

So, the idea of market as a morally free zone sets out the ideal horizon. Short of a morally free zone, we should resort to morality:

> But this is not to denigrate the value of morality, which makes possible an artificial harmony where natural harmony is not to be had. Market and morals share the non-coercive reconciliation of individual interest with mutual benefit. Where mutual benefit requires individual constraint, this reconciliation is achieved through rational agreement. (Gauthier,1986:13/14)

A necessary condition for agreement among individuals, in other words, its initial motivation, is that its outcome be mutually advantageous. Sufficient conditions entail, in turn, that the problem of achieving the mutual benefit be modelled as a bargaining problem (instead of, we now know, a repeated game or a Bayesian decision making, or more generally, a problem of decision making under some veil of ignorance, as in Rawls and Harsanyi).

Bargaining is used to select a specific outcome, given a range of mutually advantageous possibilities and a carefully specified initial bargaining position. An individual criterion for selecting outcomes in the bargaining problem is the minimax relative benefit, Gauthier suggests, which specifies the minimum expected individual surplus as a result of the bargain. This minimum defines, in turn, the willingness of the individuals to make concessions as well as the threshold below which they are no longer interested in the cooperative venture.

However, a crucial condition for his compliance with the final terms of the bargain, according to Gauthier, lies in the plausibility of his constrained maximizing behavior and in a well specified initial bargaining position (which he sets out in the form of a Locke-like proviso that protects personal and property rights).

He finally sets out to argue for an Archimedean point from the standpoint of bargaining theory rather than the usual rational choice under uncertainty perspective. He certainly claims this point to be his fifth core concept, but only addresses it at the end of his book.

The central issue that shall concern us here with Gauthier is his idea of a constrained maximizer, to whom he devotes a great amount of justificatory arguments. It seems, however, that the

assumption of constrained-maximizing behavior shares the basic shortcomings of similar behavioral assumptions by authors who deal with the emergence of political order from a collective action standpoint. It draws heavily on the capacity of actors to confidently and correctly identify other would-be cooperators and join them in the cooperative venture.

Of course this issue brings back Olson's concerns as to the visibility of free-riding. Insofar as rational agents will only cooperate conditionally on others' cooperation, which is the point of constrained maximization (agent's willingness to restrain their narrow self-interest because of the prospects of the greater benefit were cooperation to be achieved), cooperation draws heavily on rational agents' capacity to identify each other's disposition to coordinate. But, on reflection is this capacity a conceivable feature of an instrumentally rational agent? Moreover, having identified true disposition to cooperate on the part of the others, why should one cooperate oneself? Is it rational? We are back to square one, or so it seems.

5. Identitarian Collective Action

Let us now move to what we call 'identitarian' collective action, the meaning of which will emerge from some cases considered below.

According to a number of authors, self-interest is an inadequate motivation to explain social interaction. Were we to relax it some interesting possibilities might arise that admit of cooperation among individuals. I shall briefly refer to Sen (1985b), Pizzorno (1986) and Elster's (1985) contributions.

Sen argues that the Prisoner's Dilemma reflects the problem of divergence between equilibrium points and Pareto-optimal points, which by definition coincide in the model of perfectly competitive markets. Usual ways out, he goes on to state, involve the use of repeated games and lead to multiple equilibria unless the behavioral structure is altered in some significant way. Also, relaxing the assumption of common knowledge helps to escape the dilemma which is a route some authors have taken. But, he argues, this latter seems to be an odd way out for 'in order to achieve

rational cooperation, it becomes necessary to know 'less'!'.[27] His own attempt concentrates on relaxing the behavioral assumption of 'goal-priority' or '**self-goal**' choice entailed in standard game theoretical approach.

Indeed, Sen's proposal implies the rejection of rational economic man:

> The conception of the individual as a very 'private' person - unconcerned about the rest of the world - has been seen, in my judgment rightly, as both empirically unrealistic and theoretically misleading. (Sen,1985b:346)

He recognizes three types of 'privateness' which are implicit in the behavioral assumptions of game theory: (i) self-centered welfare - a person's welfare depends only on his or her consumption (and, in particular, it does not involve any sympathy or antipathy toward others); (ii) self-welfare goal - a person's only goal is to maximize his or her welfare, or - given uncertainty - the expected value of that welfare (and, in particular, it does not involve directly attaching importance to the welfare of others); (iii) self-goal choice - each choice a person makes is guided immediately by the pursuit of his own goal (and in particular, it is not restrained by the recognition of other people's pursuit of their goals).[28]

As we move from (i) to (iii) weaker senses of privateness are suggested. In (iii) sympathy may be included and it refers to an emotional link between the self and the fate of some others. Thus relaxing condition (iii) would require breaking the tight link between individual welfare (with or without sympathy) and the choice of action, 'e.g. acting to remove some misery even though one personally does not suffer from it'. This involves what Sen calls, after Bernard Williams, 'commitment':

> Commitment can also involve violation of self-goal choice, since the departure may possibly arise from self-imposed restrictions on the pursuit of one's own goals (in favour of, say, following particular rules of conduct). (ibidem:348)

---

[27] Sen, 1985, p.344.

[28] Sen, 1985, p.347.

To follow particular rules of conduct may well be a matter of identity, that is, of how one sees oneself, Sen argues. It may be crucial to the way we view our welfare, goals, or behavioral obligations, he also suggests. In particular,

> the pursuit of private goals may well be compromised by the consideration of the goals of others in the group with whom the person has some sense of identity. (ibidem:348)

Conceding that possibility, we may consider that relaxing self-goal choice may resolve the Prisoner's Dilemma, for it makes room for an alternative pattern of behavior toward others.

Pizzorno (1986) suggests something similar, when, in dealing with odd behavior by some people, he argues for a **non-rational choice** explanation. In certain cases, he maintains, there is something that precludes rational calculation or the self-interest premise for, in a sense, they involve aspects that are somehow antecedent to one's choice and accordingly specify the kind of self a person is. Some cases of collective action displays this peculiarity: people more than adhere to a cause; they might happen to **identify** with it, in the strong sense of having their identities defined by this particular collective identity (and, in this sense lack the necessary detachment which is the condition for any calculation to be performed). Instead of talking about a well structured and unified self we should talk about different selves according to the different collective identities a person may identify with.

Elster (1985) also argues for the consideration of **non-selfish** motivations as a way out to collective action problems. In particular, he refers to two classes of non-selfish actions, namely, those outcome-oriented and those process-oriented. In the former class fall altruistic and moral motivation. Altruism, according to him, is a psychological attitude of concern with others' pleasures whilst morality is a motivation that reflects an impersonal evaluation. Morality, in turn may be either a duty-motivation, like the Kantian categorical imperative, or a desire to maximize average or even minimum welfare.

In process-oriented cases, on the other hand, people are not guided by moral norms but by social norms. Social norms have to do with the way other people see one's self, and how this image affects one's self-image. So, the primary consideration when taking an action, when a social norm

is at work, is to see whether or not the action matches the self-image one possesses, one's question being: Am I the kind of person that does this?

Elster, thus, argues that 'my self-image is not a benefit: it is what defines what counts as a benefit'. much in line with the previous approaches we have referred to. However, contrasting with these latter, Elster warns about the issue of 'framing' that may be related to social norms. Social norms are quite sensitive to framing, in a way that may render them quite corruptible:

> even the most fundamental norms may lose their hold on behavior almost overnight. In the case of norms of cooperation, this is facilitated because of the ambiguity of what constitutes cooperative behavior in any given case. Abstention from voting that is forbidden by the norms of duty may turn into an act of civic duty if redefined as active abstention in protest against the system. By switching allegiance from a smaller to a larger group, cooperation and noncooperation may take on new meanings. But to know that such phenomena occur is not know when they do. Until we have a firmer understanding of such gestalt changes in normative behavior, the study of collective action may not be able to go much beyond 'thick description'! (Elster,1985:154)

Two comments apply here. Firstly, the issue of framing may endanger Tsebelis's account in the sense that we may see cooperation showing up under situations carefully theoretically reconstructed so as to look like a cooperative interaction. So, in a sense, a 'social norms' account of cooperation, or an identitarian account of it, by **presupposing** cooperation cannot get rid of the same kind of frailty some authors showed to have also affected strong conceptions of self-interest, namely, to remain essentially tautological. An alternative way, suggested by Elster, is that to shift from the deductive and normative pretence which theories of collective action often have, to an inductive 'thick description' where wrong hypotheses might be falsified.

Secondly, an identitarian general approach runs the risk of justifying (perhaps unjustifiably) some degrees of particularization of the social dimension, on grounds of identity with particular groups. So, on the one hand, bare self-interest is too unreliable a basis for constructing social cooperation, whilst on the other hand the symmetrical behavioral assumption of, say, 'identity', is no more so. 'Identity' has supported factions in history.

6. Concluding Remarks

Although not a thorough conclusion, two additional remarks are worth making. First, the rational choice translation of self-interest in terms of equilibrium-seeking behavior along with the assumption that social benefits or efficient coordination thereby result is flawed, in view of the arguments the collective action literature has been adding to the earlier canonical version of this problem. From Olson, we have a logical argument severing unconditional individual rationality from collective rationality; from Olson's opponents, we have additional conditional reasons for this logic.

Secondly, taking the individual as a somewhat enlarged entity, capable of commitment, besides or rather than interest, may open an avenue to solving collective action problems. Groups, accordingly may evade the logic of collective action to the extent that their members are not primarily self-interested. Should we welcome the news? It does not seem that we should.

On balance, one may arguably think that we need a public space but conclude that self-interest is not a very enabling basis, except, perhaps negatively as a critical standpoint we stick to, from which we might cultivate suspicions about self-proclaimed public discourse. Our attempts to address this problem indirectly from self-interest alone - that is the problem of having individuals act collectively in furtherance of a common end to make everyone better off - could not pass Olson's test. However, attempts to address it directly (self-interest as, broadly speaking, group 'identity') do not seem to allay our justifiable fears, which have been cultivated all the way from early modernity, for at least we have experienced some of its bloody results. As Hardin has recently said:

> self-interest can often successfully be matched with group- interest. And when it is, it is often appalling. The world might be a far less bloody place, and less ugly in many other ways, if many groups failed in relevant moments. (Hardin,1995:5)

Implicitly recalling Hume, he goes on to explain that this is bad because this is 'typically applicable to a group whose benefits comes from the suppression of another group's interest.'[29]

---

[29] Hardin, 1995, p.5.

Collective action with, in our terms, 'identity' may lead to factions, and with bare self-interest may not succeed at all.

Different grounds for the social order might be produced by behavioral assumptions other than self-interest and group-identity. In this direction, individuals' cooperative willingness may be rooted in a moral motivation, an assumption that is taken up in Part II, specially in the chapters on Harsanyi and Rawls. Before that, however, we shall concentrate in the next chapter on questions of strategic interaction. A natural reason for doing so is that strategic rationality or game theoretical reasoning underlies much of the collective action literature and we may gain some insight by looking into well established problems in this framework. Another less natural motivation is that the analysis of strategic rationality highlights the problematic **epistemic** nature of social interaction, leaving behind the behavioral assumption of cooperative willingness which got in the way of the episodes of public-goods that we have gone through in this chapter.

In the next chapter, then, I will be thinking of social interaction as an environment where knowledge is produced and consumed, individually and collectively. I intend to open up this issue and uncover within the limit of our present possibilities the extent to which unargued normative convictions rely on at least disputable epistemic assumptions. This move, in turn, shall open the path to Hayekian social philosophy, which will be worked out later in Part III.

Chapter 3

ON GAMES AND PUZZLES: SELF-INTEREST AS STRATEGIC RATIONALITY?
(the game-theoretical approach to interaction)

1. Introduction

In this chapter I discuss the game-theoretical image of social interaction. The view shall be taken here that the game-theoretical vision of social interaction portrays it as **strategic interaction** among rational people, that is, an interaction among utility-maximizers whose choices are interdependent. Following this vision, game theory needs to work out a particular notion of rationality, as **strategic rationality** that is sensitive to the various contingencies that typically surround interdependent choices. Of course, these contingencies reflect the fact that one's outcome depends also on the courses of actions possibly chosen by relevant others, at least by the player with whom one is interacting. Since it would be hard to maintain that we have an obvious and unequivocal access to each other's plans, in examining strategic interaction we shall find ourselves in the realm of the 'two'.

Towards the end of understanding game theory's view of social interaction, I am going to elaborate, conceptually and non-technically, on the idea of strategic interaction starting from its formulation by von Neumann and Morgenstern, and Anderson and Moore. The provocative and original insight contained in their vision of a complex interdependence will be emphasized, in particular their departure from the then fashionable behaviorist approach to the social phenomena. Besides, aspects are also discussed that relate to the notion of strategic rationality in that this is supposed to solve the strategic interaction in terms of individual rationality and an equilibrium notion. In particular, focus shall be put on the knowledge conditions often attributed to strategic agents in that these conditions seem to perform a crucial role in the resolution of the **social** interaction in terms of the rationality of the **individuals**. In this context, it turns out to be relevant to learn how far individuals are able to anticipate each other's intended actions.

On the way to approaching the meaning of 'strategic interaction', our steps will be as follows. In section 2, two episodes are selected from the early history of the theory of games, namely, the

intellectual collaboration of von Neumann and Morgenstern and Moore and Anderson's view of the superior prospects of the nascent game theory as compared to the then existing alternative accounts of human interaction. These 'historical' episodes are there for two reasons: first, to provide a perspective of the theory of games as an attempt to formalize an original and compelling way of conceiving the social phenomena. and secondly, to give an indication of some of the relevant tensions that the theory has so far been trying to case as a consequence of its twofold approach, represented in the search both for relevance (faithfulness to complexity) and determinacy (solution concepts).

It seems that the tensions that come out are to a large extent natural outcomes of the novelty game theory itself was introducing in standard scientific views regarding social interaction at the time it appeared. To begin with, game theory (GT, henceforth) was proposing a different representation of a 'rational' decision from the traditional behavioristic one, as a consequence of its peculiar view of the social phenomena. I shall quite arbitrarily represent these distinct representations as follows:

(1) Behaviorist model: $\mathbf{s} \rightarrow \mathbf{o}$

(2) GT model: $\mathbf{s} \leftrightarrow \mathbf{o}$, or else $\mathbf{s} \leftrightarrow \mathbf{s(o)}$;

where **s** is the subject, the reasoning actor, and **o** is the object of his reasoning, which in human interaction in general is another **s** and hence may be represented as **s(o)**. The arrow then indicates the expectation of an action or the construction of a belief concerning the recipient of it. Under representation (1) the object is external to the subject's undertakings, whereas under (2) the object is somehow affected by the subject's concerns.

At the same time, however, the ultimate purposes of GT are to build up a deductive theory capable of finding out the big S whose task it is to solve the system $\mathbf{s} \leftrightarrow \mathbf{(o)}$:

$S \rightarrow s \leftrightarrow (o)$ [1]

So, the point is that while GT acknowledges the existence of a modified object, in our terms the complex object $s \leftrightarrow (o)$, it also devotes itself to the construction of the theoretical external viewpoint from which a solution might emerge. In this context, some of the tensions appear to be but the fairly inevitable unfolding of GT's twofold scientific program, as I hope eventually to indicate.

On the way to approaching the meaning of 'strategic rationality', then, section 3 undertakes to explore some of the knowledge conditions of GT and their recent developments. I do this because the knowledge issue seems to play a crucial role in the theory, given its peculiar view of interaction. It becomes very important to set out what individuals know about their decision problem, and, particularly, what they know about each other's choices. In particular, in standard game-theoretical approaches the requirement that the theory of games be somehow common knowledge among the players in a strategic situation seems to be rather crucial for a determinate solution to arise (as well as theoretically demanding). So, the common knowledge assumption (CK) shall be firstly presented within the intellectual context of its birth in connection with the notion of conventions, and afterwards is discussed in connection with the solution to non-cooperative games.

Although strategic reasoning is one way of reasoning to a solution in strategic interactions, it may also engender an uncertainty of its own, a 'strategic uncertainty' (after Johnson, 1993). The assumption of common knowledge is normally thought to be of help in reducing the uncertainty that may follow strategic thinking. Different informational levels may be supposed of the players, though. As we move from the standard Nash solution to the Bayesian approach our attention is directed at weaker informational structures and stronger rationality assumptions, where the requirement of complete information is assumed to be too restrictive and not really necessary in order for rational individuals to reconcile their plans. At this point all that is needed towards that

[1] The symbolic representation $s \rightarrow o$, $s \leftrightarrow o$, and $s \leftrightarrow s(o)$ herein adopted is 'stolen' from Santos (1990) (who worked it out in a different context) and freely distorted.

end is to suppose, as Aumann (1987) does, that people are Bayes-rational and that this fact is common knowledge.

This latter solution is examined in relative depth, in section 4, as it perceives strategic interaction in a rather radical fashion and tries to address the 'strategic uncertainty'. The vision of interdependence that seems to be implicit in this solution may be provisionally described as follows:

$$S \rightarrow s''(o) \leftrightarrow s'(o) \leftrightarrow s(o);$$

where the big S and the whole sequence of small s's, according to my interpretation, end up ironically by restating the traditional behaviorist model $s \rightarrow o$ and its epistemological shortcomings which have already been remarked on in the earlier years of the theory of games.[2]

A fifth section presents, in quite a cursory fashion, three alternative approaches to Aumann, which undertake to examine necessary supplementation to strategic rationality in its comprehensive form.

In its final comments, this chapter considers that a deductive theory of rationality of the kind displayed here needs be supplemented by external elements which, in turn, cannot get rid of context. That is to say, cannot be endogenously deduced, a suggestion that is in line with the alternatives examined in the previous section. In addition, this approach hints that to work out this possibility we would be required to face the rather philosophical question: what can we possibly know? To pose the question in this manner means that the problem of knowledge is not supposed to be just a problem of information, and to solve it is not only a matter of maximization under uncertainty. It is rather the question of the place we occupy in a quite autonomous structure itself made up of knowledge, and the complex relations thus generated among the parts to it and the whole. This is the subject of Part III.

---

[2] Actually, the description that we shall provide in section 4 shows a big S conjecturing on different s(o)'s, that is, different action-reaction settings of strategic interaction.

2. Strategic Interaction

1st approach: The 'we-action'

Hardin (1988) claims that game theory is the best understanding we presently have of strategic interaction. This strategic interaction, in turn, is something that implies that, in trying to achieve some end in society, 'what **we** do' turns out to be much more important than 'what **I** do". And this fact seems to have, according to him, important moral implications, at least of one kind: that one cannot conclude from a choice of strategy whether it is right or wrong, for it alone does not imply a particular outcome, but only restricts the range of possible outcomes. So, he concludes, if we are utilitarians we should give up any precise sense of determinacy in our moral theory. But perhaps even if we are not utilitarians and still retain the image of interdependence, this implication may have something to tell our normative or positive theories.

Leaving aside, for the moment, the 'consequentialist' (or, utilitarian for that matter) consequences of this attitude towards social interaction, we may be interested in seeing in what sense game theory is said to underpin the intuition that social interaction is strategic in kind, and to what extent it offers prospects for us to limit the apparent indeterminacy involved in our understanding of strategic interaction. To shed light on these issues it is worth looking briefly at the circumstances around the collaboration of von Neumann and Morgenstern. I will, in the following, draw on Mirowski (1992) and Leonard's (1995) reports of the main elements of this collaboration.

2nd approach: The von Neumann-Morgenstern problem

> *"Whenever the outcome of an individual's action depends as much on the actions of the others with whom he interacts as on his actions (and individuals are aware of this dependence), their activities display a strategic component and thus fall within the vNM (1944) definition of a game".*[3]

In the twenties, von Neumann was highly engaged in two different projects, one being a thoroughly formal and comprehensive axiomatical project in mathematics (the so-called Hilbert's program) and the other, the quantum mechanics research program. The first project, it seems, was to be defeated, at least in its original form, by Gödel's incompleteness proof in 1930, which is

[3] Bicchieri, 1993, p.8.

known to have pointed out the endogenous generation of inconsistencies in complete logical systems (either a logical system is complete or consistent, it cannot be both).[4]

Von Neumann, after Gödel's proof, retreated from his initial position inside the mathematical field In contrast, the second project, quantum mechanics, was destined to achieve great success and had an extension in von Neumann's mathematical work on games. How did this come about?

To be sure, von Neumann's work on games was also a reaction to the failure of Hilbert's meta-mathematical project, if not in pure mathematics then in the field of a general theory of rationality, because it was also part of the original program to extend mathematical formalism to as many areas of human understanding as possible. This mathematical formalism, in turn, meant basically that 'the question of truth [was] then shifted into the question of consistency.[5]

In any case, games were already an independent part of von Neumann's research agenda, which approached his general interest in the widespread application of mathematical formalism, in line with one particular task of Hilbert's project. On the other hand, his contribution to the new physics program encompassed the development of new mathematical tools. The interesting question arises, then, of how von Neumann's striving for a general theory of rationality was to meet his parallel ongoing research on physics.

At first, we should be clear about the nature of von Neumann's efforts in physics. He was actually working on the mathematical foundations for a new physics, those which would work out the stochastic nature of the physical world courtesy of a peculiar effect that mere observation was supposed to produce on physical particles. Rephrasing it, one highly suggestive element in the quantum project from von Neumann's perspective was the introduction of stochastic relations at the most basic levels of physical postulates, namely, in the very system observer-observed.

---

[4] According to von Neumann's personal assessment: "The very concept of 'absolute' mathematical rigor is not immutable. The variability of the concept of rigor shows that something else besides mathematical abstraction must enter into the makeup of mathematics..." (cited in Mirowski,1992:122, quoted from von Neumman).

[5] Leonard, 1995, p.733, after Weyl.

Developments in set theory and combinatorial methods were needed to frame the seemingly complex phenomenon at hand.

In addition, von Neumann extended his attention to the social world as well, as he became convinced that both social and natural worlds shared the characteristic of enabling a complex relationship to arise between the observer and the observed, in the pervasive expectation that the observer will affect the object of his examination. Von Neumann began to see games as a good representation of human interaction (recall the model **s ↔ o**). It is interesting to remark here that von Neumann also adopted the methodological monism characteristic of theoreticians of behaviorist persuasion albeit with opposite visions of the physical and social worlds.

Furthermore, von Neumann's recognition of a subjective element in the core of scientifically defined systems may be seen in the special importance probability began to assume in his works.[6] There Probability was expressing not only an uncertainty **about** the observed, but also pointing to a plausible **influence** of the observer on the observed, capable of generating further uncertain relations, which the observer himself was somehow to grasp. This development added to his work on the theory of games, as witnessed by the parallel development of measures of **expected** values of outcomes and the concept of mixed strategies, which were supposed to control the influence of chance not only on the rules of the game (through the introduction of measures of expected value) but also on the behavior of players (which was to be grasped by the idea of a mixed strategy).

Thus, the need for a new measurement tool was reflecting a change in the measurable itself: from von Neumann's research on parlor games, he became convinced that nearly 'any event could be looked at as a game of strategy provided one looked at the effect it had on the participants'.[7]

So, summing up, the development of a theory of games represented von Neumann's personal efforts towards achieving a theory of rationality whose model is a game, a situation in which any

---

[6] Mirowski, 1992.

[7] Leonard, 1995.

action will be conditioned on the reaction it releases. His question being: is there a rational way of going about this situation?

Of course, a very important target of the game-theoretical project was the discovery of definite rules and solutions to games, not so far from von Neumann's early axiomatical purposes. But, in another sense, the theory of games was still to retain, and cope with, the uncertainty element that the physical project had injected in scientific inquiry in general, and in von Neumann's mind, in particular. Thus, the concepts of expected value and mixed strategies, with their stochastic meaning, provided the essential elements for the proof of a fundamental theorem within the theory of games, the minimax theorem.

However, von Neumann's insistence on a determinate though not wholly isolated system led him to propose a quite fragile solution to the 'intersubjectivity' problem that is likely to arise whenever different descriptions of the same physical event are offered by two different observers: he denied the possibility of two divergent 'consciousnesses' and proposed that the first observer should serve as a observing tool for the second. In the same way, nor could he accept the existence of conflicting rationalities in game theory.[8] It seems in this latter case that von Neumann believed that one could deduct or discount, through a suitable solution concept, one's influence on the observed object.

The question now remains how economics entered game theory's early concerns. Though it is reported that von Neumann had already made an important contribution to the economic debate at the time he met Morgenstern in Princeton (1938), it was without doubt the collaboration with Morgenstern that provided von Neumann's game-theoretical technique with a fertile field on which to expand, namely economics, or, economics as viewed by an Austrian economist.

---

[8] Mirowski, 1992.

In the early thirties, Morgenstern, following the steps of the Austrian school, felt uneasy with the current economic explanations and equilibrium concepts, not well suited, according to him, to our world of generalized interdependence and limited foresight.[9]

He too, like von Neumann, was trying to find a better physical analogy than classical mechanics to apply to economics. Interdependence of choices as a central trait of economic life indicates, he thought, the use of a notion of 'subjective' rationality in economic theory (where individuals are thought of as taking the plans of others into account when deciding on their own plans), in order to capture the strategic relations among people. A better meaning was needed that would substitute a lively representation of economic interdependence for the 'auctioneer' and the 'general equilibrium' mechanical images taken out of the Walrasian world. Interdependence (in this particular 'lively' light) and the problems of prediction that inevitably follow called for a better equilibrium concept than the received one (specially Hick's) which assumed perfect foresight. The reason was that:

> social sciences have the peculiarity of being able to affect their object of study. The prediction of the astronomer can have no effect on the subsequent movement of the stars, but that of the economist **can** change economic events. (cited in Leonard, 1995:741, quoted from Morgenstern)

Thus, our concepts of minima and maxima were to be replaced with an alternative one, able to capture (the strategic nature of) the interdependence of the agents' decisions, like the minimax. In addition, our idea of 'consistency of plans' was also to be enriched. Instead of thinking of economic agents as isolated decision makers, constrained only by unrestricted beliefs or maxims, as in parametric choices, the more likely decision setting would have us frame their decisions as constrained by restricted maxims, those which make their behavior contingent on what they assume others' behavior to be. The question of how to formalize these issues gave rise to von Neumann-Morgenstern's collaboration.

---

[9] He remarked on something he often considered to be overlooked by conventional general equilibrium analysis, that: "always there is exhibited an endless chain of reciprocally conjectural reactions and counter-reactions. This chain can never be broken by an act of knowledge but always through an arbitrary act - a resolution ... Unlimited foresight and economic equilibrium are thus irreconcilable with one another." (cited in Mirowski,1992:129fn, quoted from Morgenstern)

Again, the new research program was somehow to harbor the paradoxical aspects of uncertainty and unpredictability (which were ordinary characteristics of human interaction), with mathematical technique, within a new equilibrium framework. For von Neumann this represented the commitment to achieve a theory of (strategic) rationality that could be publicly known without this 'publicity' causing anyone to change their choices after learning it. The uncertainty aspects involved in interaction could be dealt with without any fear of infinite regress, for the theory of games would advise players to play in the uniquely rational way. For Morgenstern, the new research agenda was to provide a technical expression, through suitable equilibrium concepts, of the uncertainty and unpredictability of the social environment, in general, and economic activities, in particular. These solution concepts were expected to successfully **discount** influence of chance on the rules of the games and behavior of players.

3rd approach: The Moore-Anderson problem

> *[a game has the capacity to] "change suddenly on us with a view to thwarting our efforts at a solution..."*[10]

A similar uneasiness, this time with current **sociological** explanations of human interaction, gave rise to a parallel vindication of a game-theoretical approach to the matter. Thus, a simpler and clearer idea of some of the initial prospects of the theory of games can be seen in Moore-Anderson (1962). They proposed the games-approach as an alternative to the method that was prevalent in the social sciences at their time, namely, behaviorism. Behaviorism, according to Moore-Anderson, approached social phenomenon as if it were a 'puzzle' to be deciphered. In their criticism to this view, they contrasted the **games** representation of the social phenomena with that of **puzzles**. While the games representation describes a situation of endogenous uncertainty, in the sense that the very attempts at a solution should be seen as part of the problem to be solved, puzzles, in turn, miss the point of social interaction for puzzles cannot "change suddenly on us with a view to thwarting our efforts at a solution; (...) no social interaction takes place between the puzzle and the solver."[11] So, a puzzle may be represented by $s \rightarrow o$, while a game can be either $s \leftrightarrow o$, or $s \leftrightarrow s(o)$.

---

[10] Moore and Anderson, 1962, p.413.

[11] Moore and Anderson, 1962, p.413.

Moreover, having a similar intuition as von Neumann and Morgenstern (vNM, henceforth), Moore-Anderson nonetheless appear to go even further to the point of suggesting a nearly causal claim as to how the system observer-observed may be able to generate unexpected relations and ensuing problems for theoretical attempts at grasping them. And here we may identify a first disanalogy with vNM's view. That is, it is not only a matter of taking stock of the fact that one's observation affects what is observed, a condition that would suffice as far as a physical observation is concerned, but also that anyone might **intentionally** change the conditions of everyone including one's own observation, as in the social world. And this might be done for a number of different reasons not always clearly identifiable. Intentionality in this rather problematic sense may complicate the horizon of prediction thus rendering randomization and the theoretical strategy of 'discounting' uncertainty less attractive. So it is that the distinction between $\mathbf{s} \leftrightarrow \mathbf{o}$ and $\mathbf{s} \leftrightarrow \mathbf{s(o)}$ may become relevant.

Still, like von Neumann, Moore-Anderson spelled out some reasons against the application of descriptive mathematical methods in relation to the puzzle model of behavioral social sciences. Following vN&M, they looked for more suitable analogies between the natural and the social realms and theories. In this way, they invoked the Heisenberg Uncertainty Principle, which ruled out the possibility of a totally isolated system, that is, "even making a physical observation jiggles things a little bit - the 'totally isolated system' is affected by being observed."[12] Nevertheless, Moore-Anderson also stressed the fact that unlike the physical domain, theories about the social domain, be it either market behavior or social action, are members of it and may exert, indeed cannot help exerting, an influence on it. Actually, they stated, decent theories are expected to influence people's conduct. Therefore, they concluded in a different line from vN&M, our theories may be self-defeating in that "our predictions of human behavior might go wrong if we tell our subjects all we know".[13]

So, at this point, Moore-Anderson clearly distance themselves from vN&M who believed that, being faced with the only rational way of playing a game, as that provided by the prescriptions of

---

[12] Moore and Anderson, 1962, p.418.

[13] Moore and Anderson, 1962, p.419.

the theory, people shall adopt it. In any event, Moore-Anderson conclude, better the (incomplete) theories which are able to encompass the capacity of individual agency of self-consciously following and formulating some social rules, and do not assume, as behaviorism does, that people just happen to act on externally given problems. In this case, they go on to argue, we should give up a mathematical theory of human behavior and stick to a "theory of cultural objects which people have invented and (or so we claim) used to help shape their behavior."[14] Among these objects they include the folk-models of game-theory. Unlike behavioral and learning theories, game-theory is not primarily concerned with a psychological or biological non-contextually defined framework, but with social concepts or group sociology.[15]

Summing up, the two seemingly most interesting messages conveyed by Moore-Anderson's view are that first, there is something peculiar in the influence that observation might exert on the observed in the social sphere which makes it refractory to a behavioristic (non-contextually defined) treatment, for in the social sphere we are dealing with intentions, surely a hard issue. In other words, the fact that the agents are aware that they are the object of each other's conjectures may affect their behavior in quite unpredictable ways. Secondly, a theory of social interaction is a member of social interaction, hence it cannot avoid the difficulties concerning self-referential reasoning, that is, it cannot be complete, as in affecting its object it in turn becomes part of it. Therefore complete publicness of the theory seems to be logically impossible.

In conclusion, we seem to have here two divergent programs, one engaged in the search for determinate solutions to the games, the other claiming the less ambitious purpose of properly describing strategic situations and pointing to a non-deterministic and non-purely deductive approach.

The next two sections will present some attempts at determinate solutions respectively in classical and modern game theory. Underlying these views, we will find either strong informational structures or strong rationality assumptions. In either cases my conviction is that the knowledge

---

[14] Moore and Anderson, 1962, p.426.

[15] Moore and Anderson, op.cit.

issue is poorly addressed, and as a result GT fails to differentiate itself in a meaningful way from the puzzle model.

3. Strategic Rationality (i): The 'Common Knowledge' assumption

3.1. Common Knowledge and Convention

The kind of interdependence with which classical game theory is concerned seems to be the interdependence betwee one's own outcomes and the someone else's choices. However, there is something that it is willing to maintain, which is the **independent** character of the decisions taken by the individuals. The independent and conscious individual actor, that is to say, somehow has to take the extant interdependence of choices into account when privately deciding how he himself shall act with that knowledge and when aiming at a certain outcome. His choice can be made to be very simple, then: given the known rules of the game (the strategies and preference structure of the players, and payoffs) he will select a strategy which is his most preferred given the choice of the others.

The knowledge requirements, unlike the usual requirements that classical decision theory assumes, refer not only to an acquaintance with some more or less precise possibilities, say, likelihoods of some definite external events, but also to the decisions that might be taken by a symmetrically situated actor. If what I may obtain is contingent on the actions taken by my opponent I should take the trouble to try and outguess his plans. The only safe way for me to outguess his intended actions is to suppose that he, like myself, recognizes the situation we are in as a particular one, as a game with particular rules. The claim by standard GT is that if we want theoretically to determine a solution to this decision problem we should then note that the relevant knowledge requirement is the 'common knowledge assumption' or CK. We should assume that the rules of the game are known by each player, and that they are known to be known by them, and known to be known to be known, and so on *ad infinitum*.[16] More generally, an event **e** is CK of order **n** whenever (everyone knows that)^{n} **e** is true.

---

[16] For a technical definition, the seminal reference is Aumann (1976). Many games may require less than infinite CK for a solution to emerge, such as those where dominant strategies obtain, like the Prisoner's Dilemma.

How could we make sense of this logical condition? The common knowledge concept was first worked out by the philosopher David Lewis (1969) in the context of his discussion of conventions which afterwards gave rise to developments in the literature of conventional coordination-equilibria. An example of a coordination game that may reach a conventional solution is provided by the 'game' in which one is involved when, while talking to a friend on the telephone, the call is suddenly interrupted. What should one do? A convention might help one in finding one's way. Working out some of the main abstract features of the situation would mean framing the general traits of a conventional solution to coordination games.

The possibility of a conventional solution to coordination games, says Lewis, relies upon the stabilizing influence of an underlying assumption of conventions: common knowledge. That is, a convention is firmly at work whenever everyone conforms to it, and this obtains **because** (given that there is a general preference for conforming behavior by everyone) everyone believes that everyone conforms to it, and so on. The ideal circumstances of conventional coordination equilibria are those in which coincidence of interests among the people involved dominates over conflict of interests and this fact is also common knowledge.

In our example there is a general and commonly known preference for resuming conversation and whenever there is a convention, implicit or otherwise, that, for instance, the person who called in the first place should try again, the convention must be common knowledge for it to succeed in achieving the communication, for any doubt may lead to a coordination failure.

Besides the stabilizing force that common knowledge is supposed to inject in coordination, Keynes considered the symmetrically critical element it may introduce under the circumstances of a sudden breakdown of a convention, that is, a change in its direction.[17] This is illustrated by the overall imitation that may take place in financial markets, for example, due to a breakdown of a previous state of expectations, and which may lead to an endless 'specularity' in the form of financial speculation. It has been reported, for that matter, how prices may change abruptly due to a sudden change in the general expectations concerning stock prices, as when someone feels that the price of the stocks he possesses is going to fall and subsequently attempts to sell them.

[17] Dupuy, 1989.

Insofar as this behavior is imitated by others, prices may fall rapidly and confirm the original expectations that indeed led to this state of affairs. According to Keynes, the uncertainty that characterizes many economic decisions and above all those in financial markets makes imitation genuinely rational way for an individual to go about things. People imitate those they judge to know more than themselves. Since nobody actually **knows** anything but just conforms to what others are doing and since everybody conforms to the new convention that prices are going to fall and everybody is known to conform to it and so on, the best one can do is to try to get rid of one's stocks as soon as possible and avoid the inevitable disaster that will nonetheless befall everyone.[18]

Though this process can in one way or another be curtailed, it shows that the **arbitrary** nature of conventions enables them to give rise to instability through the very same 'specular' mechanism of common knowledge whereby it operates, and of which 'unanchored' imitation is a problematic specimen. By 'unanchored' imitation it is meant imitation not grounded in any good evidence, of real benefits. The expectational mechanism underlying our actions may lead us to conform **any** convention, a stabilizing (in Lewis's sense) as well as a critical one (in Keynes's sense). So, CK is something that can be found either beneath coordination or coordination failure.

Dupuy (1989) examines the possible ways of reconciling this common knowledge assumption with the 'finitude of human mind' so that the infinite process of the kind 'I know **x**, he knows **x**, I know that he knows **x**, he knows that I know **x**, and so on' could be accommodated within the limits of human mind, in terms of possible human practices. The most noteworthy are the interpretations of common knowledge as 'common sense', as some conventional points on which people just happen to coordinate, and the idea of a 'bootstrap' mechanism, which he identifies with the assumption that the players in a game know the theory of games. In this latter case, players are relieved from engaging in the infinite chain of reasoning entailed in the logical assumption of common knowledge as long as they adhere to the prescriptions given by the theory, which would provide them with rationally deduced focal points, so to speak, or the theory's solution concepts. This latter interpretation of common knowledge implies that every player is able to the others' decision according to the canons provided by the theory. That is, each player is able to grasp the

[18] See Keynes (1973) and Orléan (1989).

other's method of choice, for this is the same he himself follows, that is to say, the prescriptions given by the theory.

The suggestion of interpreting common knowledge as common sense is not undertaken by standard game theory. In any case, towards the end of this chapter a brief mention shall be made of recent efforts at accommodating common sense within a game-theoretic framework (Scharpf, 1990). In contrast, the second suggestion, that of taking common knowledge as the assumption that players in a game know the theory of games, has pervaded standard GT. It aims to solve the problem of the infinite mental process involved in CK in a way that enables the convergence of people's plans.

3.2. Non-cooperative games and the standard Nash concept

Conventions presuppose common knowledge, but the converse is not true, or so standard game theory has claimed. In addition, conventions are arbitrary in origin and quite ambiguous as a guide since they may generate good or bad equilibria. Common knowledge, then, cannot be relied upon to guarantee social interaction as it is framed within the context of conventional equilibria.

However, CK may function in another way when related to the idea that the implicitly infinite mental process supposedly held by players may converge to a predictable rationally deducible point, an equilibrium of a sorts. Invoking the principle of parsimony, let us assume that we want to frame social interaction as the result of the interaction among rational agents alone, and we do not want to make it rely upon any extra substantive assumptions concerning the desirability of interaction or agreement itself (as entailed in Lewis's analysis). Suppose, then, that people are likely to have an interest in cooperation only insofar as this interest does not conflict with individual rationality, that is, maximization of one's utility. For this reason, it is time to examine the case of non-cooperative games.

These are games that typically represent a mixing of conflictual and cooperational aspects. They generally involve some common purpose (either to achieve some positive aim, as in the Prisoner's Dilemma, or to avoid a decidedly negative outcome, as in the Chicken game) which nonetheless plays a non-dominating role in their solution. Now, my aim is to examine how one particular

solution concept was elaborated on which supplied the assumption of CK with a 'focal point' of sorts.

So, the basic non-cooperative non-zero sum games reached a solution with Nash's (1951) development of an equilibrium concept for these games. According to Kreps (1987), since then the Nash equilibrium has been deemed to be a necessary condition for an agreement among rational people to be self-enforcing. That is, if an agreement is to be achieved among rational players under the non-existent institutional structure of a non-cooperative game, it has to pass Nash's test.

So, let us turn to the definition of a Nash equilibrium, as proposed by Kreps:

> A Nash equilibrium (in pure strategies) is a strategy profile **s** such that no single player, by changing his own part of **s**, can obtain higher utility if the others stick to their parts. (Kreps, 1987: 584)

Now, the notion of equilibrium broadly implies the idea that nobody is able to improve his situation by deviating from the equilibrium strategy.

> Generally speaking, an equilibrium is a situation in which each individual is doing his best, **given the actions** undertaken by others and **the institutional constraints** he faces. (Bicchieri, 1993:1, my emphasis)

The emphasis in the above quotation, however, intends to underline the aspect of parametrization concerning the equilibrium play by a player, that is, he must take some aspects of his choice as parametric in order to identify his equilibrium strategy. As the choice by others is a rather problematic element, the parametrization of it turns out to be the crucial question an equilibrium concept must address.

In the context of the Nash equilibrium, it may be argued that underlying the choice of a Nash strategy there is the condition that the player is able to anticipate the other's move and select his best reply to that move. Insofar as each other player is symmetrically able to anticipate this, he is to ratify his opponent's reasoning by choosing the action that the opponent had outguessed

because this is anyway his best reply to the intended play of the other, and so on.[19] Therefore, rationality (utility maximizing behavior) on the part of the players, and CK thereof, is required for this solution to arise. Note that the choice of the optimal strategy presumes that players' beliefs are rational in the sense that these are beliefs concerning each other's choices that are mutually coordinated or consistent. Everyone does what is expected by everyone else.

A commentary by Heap and Varoufakis (1995) remarks, in this line, that the Nash concept indicates the strategy that is supported by beliefs which do not presume that one's opponent will make a mistake by expecting something which the other does not intend to do. Nash strategies are often referred to as self-confirming strategies, or as 'the obvious way to play'.[20] The assumption here seems to be that **internal** consistency of beliefs is sufficient to assure **mutual** consistency of beliefs. Again, this is tantamount to saying that players know the theory of games (they would have read a book of it and would have followed its advice, which is opposite to Moore-Anderson's prediction).

Now, one problematic part of the definition seems to be precisely 'if the others stick to their parts'. Why should these others be thought to do so? Is this rational? It is conceivable that a situation may arise where a doubt concerning what these others intend to do might jeopardize the self-confirming reasoning of Nash's concept so that one no longer has a strong reason to play his Nash strategy,[21] and instead is given a reason to deviate from it. This contingency seems already to have puzzled Anderson and Moore. If we admit this, we no longer have 'an obvious way to play the game'. As one comes to entertain doubts as to the extent to which one's opponent will stick to his Nash component, so the rationale of one's playing his component becomes itself flawed.

---

[19] Note, however, that this is the standard way in which non-cooperative GT interprets the Nash solution. However Weibull (1995), an evolutionary game-theorist, rescues Nash's unpublished Ph.D. dissertation to point out that the Nash equilibrium concept is also compatible with less than full-blown rationality as in a "population-statistical interpretation of his equilibrium concept" (p.xiii, fn). See also Weibul and Bjornestedt (1996). I am grateful to Alan Kirman for calling my attention to this point.

[20] Cf. Kreps (1990), p.404.

[21] See Aumann (1987), Basu (1990), and Tullock (1992).

Under this particular perspective, Nash's solution appears to be too strong in its underlying behavioral assumptions, since it requires implausibly strong constraints on the conjectures held by players. That is, it only works if players' conjectures concerning each other's actions are correct. It is too strong precisely in that it is not robust (after the jargon), as any plausible deviation on the opponent's part might render one's Nash play wrong.

Another problematic aspect of the definition, as remarked by Kreps (1987), is that if it is supposed to spell out the kind of rationality which prevails in a non-cooperative game, the likelihood of cooperation among players is dramatically low. Indeed, Nash equilibria will too often share the characteristic of being inefficient, a result which is unwelcome by anyone willing to ground people's cooperation in strategic rationality. Moreover, Kreps went on to point out, the Nash solution may give rise to multiple equilibria and claim, for that matter, for a theory of equilibrium selection.[22]

Still, Nash's solution may be useless in games where it provides no equilibrium points at all. So, together with the multiple equilibria issue, Nash's concept appears rather unhelpful in this light,

---

[22] Actually, the literature regarding refinements on Nash equilibria is concerned with proposing stronger logical tests than those entailed in Nash's proposal (which basically entails iterated elimination of dominated strategies). Thus, on its way to providing more restrictive conditions that might afford both unique and compelling equilibrium points, this literature is willing to incorporate in the equilibrium strategies a measure of predictable deviations from Nash's strategy profile, thus enlarging the initial belief-structure of each player concerning each other's behavior. Kreps assesses the situation in the following way: "For normal form games, one is concerned with how strategies perform if one's opponents take actions that have 'zero' probability in the equilibrium; one has to reason about the relative likelihood of things one feels are not going to happen with any significant probability. In terms of extensive form games, this tension takes the following form: Most refinements that are based on extensive form considerations are keyed to what will happen 'out of equilibrium' - that is, at points in the game tree that will not be reached if the equilibrium is played." (Kreps,1990:418). However, refinements have been inconclusive as well, for the question (as Kreps puts it) 'how can we reason about what will happen conditional on things that are not meant to happen in the first place?' does not have a unique answer. And this may be a reason for theoreticians to move towards 'repetition'. This latter move would tend to relieve the rather strong conditions on beliefs and would give room to a learning process of sorts. Yet, according to Binmore (1993:357), the accomplishment of the standard literature dealing with repetition has been to show that 'cooperation is not necessarily irrational when the Prisoner's Dilemma is repeated an indefinite number of times'. Still, as Alan Kirman pointed out to me in personal communication, repetition could be approached in terms of a learning process under an evolutionary game-theoretical key. This view however would trespass the boundaries of this section.

for it is scarcely restrictive as a solution concept. Or else it may look quite abstruse in games where the uniquely optimal solution it proposes looks unattractive.[23],[24]

In conclusion, Nash's standard solution to non-cooperative games may be considered, on the one hand, helpless in its inability to provide unique or even determinate equilibrium points, convincing equilibria in some cases, or hope for the emergence of cooperation. On the other hand, though, it can be seen as too strong in its underlying assumption of 'rational beliefs', unambiguously pursued by rational agents, whose rationale is unique. The solution, then, seems to be twice bad heuristic, as the heroic assumption concerning beliefs that supports it - that internally consistent beliefs also are mutually consistent ones - is unable to provide determinate and unique results.

One way out of the objections raised to Nash's basic concept is Aumann's 1987 article. He proposes a more general form to the Nash equilibrium, restating its assumption of common knowledge as CK of rationality. Aumann's move implies a relief from the strong conditions on beliefs assumed by Nash, and at the same time it also implies a strengthening of Nash's rationality conditions with the adoption of **Bayesian** rationality: that is, it will accept whatsoever beliefs about beliefs provided that the former be generated or updated according to the canons of Bayesian rationality.

In order to reach Aumann's solution we should first go through Harsanyi's relaxation of the complete information structure which was entailed in the standard Nash solution.

---

[23] I have in mind, for instance, the well-known chain store paradox reported by Selten (1978).

[24] So, in terms of the three main justifications of a solution concept in GT, namely, that it be congenial to individual rationality, equilibrium (a requirement concerning coordination of actions/expectations among individuals), and efficiency (a requirement concerning the prospect for Pareto-optimality), the Nash standard concept appears rather unjustified. The hypothesis underlying the equilibrium play is that rational players' actions are supported by mutually consistent beliefs, so that in order to achieve an **equilibrium** we should add to individual rationality a further restriction on the beliefs entertained by players. And this restriction itself is not justified, not even in the name of heuristic benefits for the theory, because in many cases it cannot stop multiple equilibria arising, in which case Nash's solution does not strictly speaking prescribe any. As for the **efficiency** of Nash equilibria, it has been longing for new prospects under the research program of infinitely or indefinitely repeated games, which seem to have performed poorly for that matter.

3.3. Types and Mistakes

Harsanyi (1967/1968), on his way to dealing with informational uncertainty, developed a model of the transformation of a game of incomplete information into one of complete information (to be sure, into a game of **imperfect** information), and developed, in this regard, the concept of Bayesian (Nash) equilibrium. Its rationale is to internalize the informational uncertainty so that a player can attribute probabilities to the unknown courses of actions possibly undertaken by the other players, according to a certain number of 'types' these latter might assume. In this way, one's opponent may be labeled 'tough' or 'soft' according to his disposition to accommodate or to retaliate to one's actions. Hence, having a probability estimate on the 'type' the other player might be (or impersonate) he is able to find his best response to his opponent's strategy choice. We have then transmogrified the incomplete information as to the opponent's payoff into the information about his expected payoff given his type.

To the Nash notion, Harsanyi added the Bayesian rationale concerning the formation and revision of beliefs. So, in games with this level of uncertainty, where to a large extent one does not know with whom one is playing, the supposition is that players will assign probabilities to the possible events, and, in particular, epistemic or subjective ones, reflecting their personal beliefs as to the occurrence of those events. Moreover, without any previous experience of the occurrence of these events, players will entertain the same prior beliefs about the likelihood of any of the 'types' which are the surrogates of the real players.

Hence, after Harsanyi's transformation the payoff matrix increases in complexity due to the incorporation of probabilities in the above sense, but it is still visible. In the words of Heap and Varoufakis:

> all that is needed is that you hold common prior expectations with your opponent(...) about the likelihood of your opponent turning out to be one type of player or another and the game has become one of complete information. (Heap and Varoufakis, 1995:29)

The adoption of subjective probabilities makes room for the definition of a unique **rational** behavior even under incomplete information. Under this circumstance, the rational behavior is that consistent with the logical constraints imposed on subjective probabilities.[25]

As for the possibility that the agents make mistakes in their strategic choices, Harsanyi (1973) provides a rationale for the use of mixed strategies that takes them as reflecting one's sensitivity to the small mistakes or deviations from the equilibrium strategy profile conceivably made by others.

In relation to that, a crucial question seems to be then how to anticipate behavior that is not expected on the part of rational players? And, then, how to design a rational policy to face the possible odd behavior on the part of those with whom one has dealings?

According to Aumann (1987), uncertainty is characteristically widespread in strategic interactions. It is so pervasive as to render it nearly vacuous to affirm the independent character of personal choices (as does classical GT). In this way, the information that each person possesses is not so much important as his or her estimates on what relevant others conjecture about the information that he or she possesses. It is on the basis of these conjectures that the person's opponents will decide on what to do, and their 'decisions' in turn will affect the person's decision.

In this respect, the epigraph to Aumann's article is illuminating:

> "O wad some pow'r the giftie gie us
> To see oursels as ithers see us!"

The focus on the information possessed by the agents, in particular the requirement of complete information as well as CK thereof, which was characteristic of the classical phase of game theory, is now being replaced with an approach that emphasizes strategic uncertainty and where standard solution concepts as the Nash equilibrium and Nash Bayesian equilibrium are considered as particular cases. Types and mistakes would then be given greater room.

---

[25] These conditions may be found in Savage (1954).

4. Strategic Rationality (ii): Common Knowledge of Bayesian Rationality, after Robert Aumann

In his discussion of correlated equilibria, Aumann enlarges the uncertainty space in order to encompass every aspect of the decision problem that may concern the player including the information he himself possesses, except the rationality assumption and CK thereof. In fact, informational requirements are replaced with the assumption of Bayesian rationality.[26]

So, players are rational in the sense of being utility maximizers under an uncertain environment, which means that their choices are constrained by the consistency conditions imposed on uncertain choices. This, in turn, basically means that the players are able to impute personal or subjective probability numbers (in Savage's sense) to every possible event that pertains to their decision problem.

The set of beliefs that each player maintains as to the parameters of the game corresponds to the set of the 'states of the world', each one being a particular combination of those parameters. So, uncertainty refers to what state of the world among the many possible ones will prevail. The players then attribute probability numbers to each possible state. Each state of the world, in turn, reflects a correlated equilibrium (a non-independent one), that is, a situation where parameters are given and the equilibrium strategy profile is deducible from maximizing behavior as well as from the relevant CK.

The idea of an independent personal choice, Aumann observes, no longer makes sense as long as the strategic decision problem is modelled in this way. The reason is that the rationality of one's personal choice is in a non trivial sense contingent on the personal choice of the other player with whom one is interacting, as the 'personal choice' seems to depend not so much on one's conjectures on the other's possible actions as on one's estimate of the other's estimate of one's own choice.

---

[26] In Aumann's words: "indeed, in our treatment, the players do not in general know how others are playing. We assume only that it is common knowledge that all the players are Bayesian utility maximizers, that they are rational in the sense that each conforms to the Savage theory." (Aumann, 1987: 2)

4.1. Probabilities, Bayes's rule and Common Priors

Bayesian rationality implies more than that players are able to assign subjective probabilities to everything. It also implies that the probabilities thus assigned somehow converge to 'objective' probabilities. Aumann tries to find some support to this view in the developments of probability theory by Savage. However, as shall become clear, Savage recognizes at least three distinct interpretations of the mathematical concept of probability.

So, according to Savage (1954), the mathematical concept of probability is crucial for the definition of consistent action in the face of uncertainty. However, he goes on to remark, though the axiomatic concept of probability is fairly clear, there is much disagreement regarding the interpretation of its meaning. In particular, Savage identifies three main classes of interpretation of the mathematical concept of probability: (i) the **objectivistic** (or frequentist, as it is presently known), which understands probability as reflecting observed repetitions of an event; (ii) the **personalistic** (the view espoused by Savage himself), which sees probability as measuring the degree of confidence of an individual in the truth of a particular proposition, allowing for different such degrees among reasonable individuals faced with the same evidence; and (iii) the **necessary**, for which probability measures the extent to which one set of propositions, out of logical necessity and apart from mere opinion, confirms the truth of another.

So, having in mind this map, a Bayesian view would imply taking both the **personalistic** and the **necessary** views of probability to the extent that it accepts that probabilities are subjectively generated as in the former, but requires that they converge to an objective measure whose content is given neither by empirical frequencies (as in the objectivistic view) nor by any conceivable intersubjective agreement, so to speak, but by some prior or logical necessity (as in the necessary view).

Savage's objections to the necessary view are worth quoting:

> it is appealing to suppose that, if two individuals in the same situation, having the same tastes and supplied with the same information, act reasonably, they will act in the same way. Such agreement, belief in which amounts to a necessary (as opposed to a personalistic) view of probability, is certainly worth looking for.

> Personally, I believe that it does not correspond even roughly with reality... I do insist that, until the contrary be demonstrated, we must be prepared to find reasoning inadequate to bring about complete agreement. (...)
>
> It may be, and indeed I believe, that there is an element in decision apart from taste, about which, like taste itself, there is no disputing. (Savage,1954:7)

To be sure, Savage also contends that personal beliefs that prove wrong when confronted with a (somehow intersubjective) consistency test should be corrected. In my judgment, however, this is not to subscribe to the objectivistic, let alone necessary, views. In any case, the decision problem Savage was addressing was that of choice under uncertainty as to **external** events. We may justifiably ask whether there is any sense in talking about somehow objective probability estimates where there is interaction between agents and environment, as happens in social interaction. There, probability estimates should reflect agents' beliefs or expectations concerning the behavior of others who are also entertaining expectations concerning the actions of the former. The question naturally arises of whether is there a rational procedure whereby these beliefs are generated or revised so that they might turn out to be genuinely rational beliefs, and, thus, converge?

According to 'Bayesianismists',[27] there are two devices that may guarantee both the internal and the intersubjective consistency of our beliefs, rendering them rational in a quite strong sense. These are the Bayes's rule and the assumption of common priors (CPA).

Originally, the Bayes's rule was designed as a device for the computation of posterior probabilities.[28] Bayesianismists, however, insist on interpreting it as a mechanism of 'learning' or as an inferential process. Very often, non-Bayesianismists (but still Bayesians, like Binmore) object to the strategy of interpreting the Bayes's rule as either a learning mechanism of a sorts (Binmore, 1993) or a synthetic description of a rational mental process in terms of an inferential process (Heap and Varoufakis, 1995). According to Binmore, the Bayes's rule can hardly be seen as a genuine learning process for when the agents come to revise their beliefs in the manner

---

[27] Binmore (1993,1994) coined the term *Bayesianismists* to designate those people who use the Bayes's rule as a mechanism of 'learning'.

[28] It computes the conditional probability as a ratio of the joint probability and the prior probability.

indicated by the rule then the real learning process, in the sense of getting an understanding of one's experience, is actually over. In the suggestion of Binmore, this 'learning' process is just a passive process of absorption of unambiguous data and being so understood it overlooks the not unusual fact that our becoming aware of particular facts might have the effect of changing precisely what we want to 'learn'. Heap and Varoufakis (1995), in turn, seem to warn against the excessive simplification involved in the reduction of a rational mental process into an inferential process alone, as this provides a very partial and mechanical image of rationality. They appear to be concerned with the exclusion of the processes of judgment also connected with the idea of rationality whereby we normally regulate the very application of the rules of logical inference. In this regard, they ask: if reason is but a set of rules of inference 'can any finite set of rules contain rules for their own application to all possible circumstances?'.[29]

As for the CPA, Aumann (1976) states that with common priors people cannot 'agree to disagree' even though they started out with different information. As long as they know the past record of the occurrence of some event (the 'posteriors'), they will come to revise their original prejudices and come closer to the 'true' probabilities. Generically, the only motive that would lead rational persons to diverge in their probability estimates would be their having access to different information; in the absence of this difference they should tend to entertain common priors.

Actually, this is the so-called Harsanyi doctrine, and it goes like this:

> Differences in beliefs are explained by differences in information; therefore individuals with the same information must have the same beliefs. (Morris,1994:236)

CPA turns out to be a very important assumption in interactive environments where there is such differential information.

According to Morris (1994), though, the CPA is not as innocuous as it may seem at first glance. For one, the co-extensitivity of beliefs and information may be objected to, even without raising the issue of interpretation, by referring to the work of Savage who developed the personalistic

---

[29] Heap and Varoufakis, 1995, p.59.

notion of probability on the sound premise that with the same information people may entertain different beliefs and still be rational. Of course, as we have already remarked, Savage adds that this disagreement cannot last in the face of the consistency tests imposed by the numerical properties of probabilities on our personal estimates. Nonetheless, as Morris remarks, we are here dealing with probability estimates as to the occurrence of endogenous events, whose probability of occurrence is affected by the actions of the individuals whose behavior we want to model. In this case, the meaning of CPA becomes even more obscure, for we are dealing with probability estimates of the behavior of others (which, in turn, is affected by the beliefs we hold about them) instead of some external event, like the weather.[30]

The introduction of the CPA by Aumann seems to be targeted to guaranteeing that even in an environment of strategic uncertainty (that generated by strategic reasoning itself) and generalized ignorance as to the main parameters of the game, it is still possible to reason to an equilibrium. Recall that, in contrast to Nash's standard solution, the notion of correlated equilibria proposed by Aumann is weaker in its beliefs-scheme insofar as it does not require certain knowledge about the strategy chosen by the other player, or, in other words, it does not assume that internal consistency of beliefs leads of necessity to mutual consistency. At the same time, Aumann's equilibrium notion seems to address the kind of uncertainty that Anderson-Moore had in mind when they suggested that agents' **intentional** choices might render their coordination theoretically intractable.

As a consequence, if only we could accept the CPA while at the same time interpret the Bayes's rule in the manner indicated by Aumann, we could free the notion of rationality from its troublesome corollary of intentionality, and by so doing we would have entered into game theory's modern phase. In particular, we would have found a procedure whereby the beliefs of rational individuals come to be aligned.

---

[30] The formal problem, according to Morris, "is that assuming a common, logical, prior about endogenous events makes the logical relation (if one existed) between information and priors self-referential." (Morris,1994:236) As for the empirical content of the difference between information and priors, matters are just as difficult. Take the following example: "If you are an 'expert' on U.S. politics, I may well want to alter my beliefs, on learning your beliefs, about who will win the 1996 presidential election, even if you have already told me all the relevant 'information' in the usual sense. Your 'expert prior' is an information signal for me, so we must interpret it as information." (Morris,1994:240)

From the preceding, a question imposes itself: what is the new 'agency' in Aumann's model?

4.3. The Outside Observer

In a world of widespread uncertainty, the assumption of independent choices or intentional and conscious action seems to be useless. As stressed in the epigraph chosen by Aumann, strategically interacting agents show great concern with their images as perceived by the others with whom they are interacting. In this sense, an individual can be thought of as at least two different 'persons'; indeed, he can detach himself from his actual self and conjecture from 'outside' how many different persons he can be supposed to be. For each of these 'persons', that is to say, given some parameters and a certain information level that is attributed to him, there is an equilibrium strategy that he can 'choose' given the choices of the other(s) individual(s) with whom he is interacting. Inside each of these 'states of the world', the individual chooses his optimal strategy, that which maximizes his expected utility, and all he needs to know about the other(s) is that he (they) is (are) Bayes-rational like himself, and that this is common knowledge among them.

So, the individual from 'outside' considers how many different 'persons' he can be, according to the different beliefs the other players hold about his information level, and attributes a probability distribution to the *n* states of the world thus obtained. In each of these states, a correlated or non-independent equilibrium obtains as a corollary of the assumption of maximizing behavior and common knowledge thereof.

Now, in each state of the world, the players 'choose' in the sense of classical game theory, for given the choice of others each one selects his optimal strategy. However, which state of the world shall prevail - and accordingly what 'person' the player shall turn out to be - is ignored by players. This situation is expressed in the probability distribution attributed by the 'outsider' to these states.

The assumption of an outside observer is vindicated by Aumann in terms of a standpoint that we may take which provides a thorough perspective of the set of possible contingencies that might affect one's decision. From a theoretical point of view, the external observer perspective is

justified by Aumann as a device designed to lead with the radical ignorance that his model allows; it would allow a full specification of an all-inclusive model.[31]

In terms of our notation:

$$S \rightarrow \begin{array}{l} s_1 \leftrightarrow s_1(o) \\ s_2 \leftrightarrow s_2(o) \\ s_3 \leftrightarrow s_3(o) \\ \quad \ldots \\ s_n \leftrightarrow s_n(o)\,, \end{array}$$

where each $s \leftrightarrow s\,(o)$ is a possible state of the world; j=1,...,n is a finite number of such states.

In this context, then, uncertainty is translated into a distribution of subjective probabilities assigned by an external observer over the set of possible contingencies that surround people's decisions. In contrast to the classical Nash concept, and even to the Nash Bayesian version, Aumann's model represents a genuine expansion of the uncertainty space since it allows for a large variety of beliefs about others' beliefs.[32]

A whole series of questions emerges in reaction to Aumann's undertaking. To begin with, how can we justify the external observer viewpoint in terms of Aumann's own initial vision of strategic uncertainty? Two difficulties would then have to be addressed: the one of epistemological nature, and the other of logical nature. From the epistemological viewpoint, if there is no metaphysical assumption excluding the 'outsider' from the very system of beliefs that he tries to give an expression via his probability distribution, there is no point in overlooking his own beliefs as mere beliefs and no justification at all to exclude the beliefs of this 'meta-player' from the set of possible

---

[31] In Aumann's words: 'Player i must take the ignorance of the other players into account when deciding on his own course of action, and he cannot do this if he does not explicitly include in the model signals other than the one he knows he got. The 'outside observer' (...) is thus a surrogate for the ignorance of the system as a whole - the lack of common knowledge [of information in our sense] - of the signals received by each player." (Aumann,1987:8).

[32] Notwithstanding, as we have seen, this extension comes at a non trivial price, namely, that players' initial beliefs are deemed to be homogeneous so that we actually do not have the interaction of genuinely heterogeneous agents, holding different yet plausible beliefs, courtesy of CPA and Bayes's rule.

states of the world. However, if this is so there is no point in disregarding, in turn, the set of beliefs of the 'meta'-meta-player and those of the 'meta'-meta-meta-player and so on. But this situation would be logically unsound in the event of the set of possible states of the world being finite, for in this case it could not contain the list of the states, and also the list of the list and so on.

However, the set of states of the world might conceivably be open or incomplete, and this seems to be Aumann's suggestion when he states that his approach deliberately neglects the distinction between 'internal' and 'external' states.[33] If this is the case, we could not justifiably use probabilities, at least after Savage's canonical approach to it, for the use of probabilities is only legitimate, Savage warns, in a closed universe where you can 'look before you leap', in opposition to an open universe, where you can only cross a bridge when you reach it.[34]

In addition, we would also have to argue for the practical possibility of such capacity on the part of the players, who, although radically ignorant, are endowed with ilimited cognitive intelligence and computational abilities. If the model is heading in the right direction, the (meta)^n-player, endowed with hyperrationality will replace the behaviorist scientist who imagined the social phenomenon as thoroughly external to himself. In this way, the strategic interaction will have been solved in terms of a behaviorist experiment: the game representation would have collapsed into a puzzle representation, endogenous uncertainty will have collapsed into exogenous uncertainty, and Moore-Anderson's problem will not have not reached a genuine solution. (Thus, $\mathbf{S} \rightarrow \mathbf{O}$, where **O** stands for the range of all possible 'states of the world', as if they were thoroughly 'external', and **S** is the reasoning outside observer.)

In a less-than-perfect world, the image suggested by Aumann in the epigraph of his 1987 article could add more nuances and overtones than light: the fact that when deciding we would have to take into account not so much the information we know we have as that we suppose others have about what we know does not allow us to take these others as mirrors on which we might see our

---

[33] See Aumann (1987).

[34] See Savage (1954) and Binmore (1993).

image clearly reflected. The image of ourselves that the others give us back must be affected by the image we have of these others, and it will already be obsolete at the time we get it back. Too much 'information', or too extended a capacity of generating information seems not to be really restrictive enough a condition (in practical terms) while at the same time too restrictive (in epistemological terms). It requires that we be able to lift ourselves by our own bootstraps resorting to rationality alone, but it actually achieves the 'bootstrapping' by appealing to implicit substantive assumptions about beliefs (such as our metaphysical ability to anticipate all possible contingencies and define an unconditional and unconditioned policy to cope with them, as well as the assumption of convergence of beliefs).

A good reason for further investigation seems to be the suspicion that measuring ignorance is feasible only in the event that 'measurement' and 'ignorance' are sufficiently independent. If not, we should give up measuring and try something else.

5. Three Alternatives: institutions, selection and free will

We have already mentioned Scharpf's objection to the foregoing argumentative strategy, against, that is, strong rationality assumptions. Here is a good occasion to turn to it. Sharpf prefers to stick to the classical informational requirements rather than to unlimited computational abilities of human beings.

In addressing the 'real critique', he argues that many mechanisms may be invoked to enhance the informational basis of social interaction, and thus to increase the degrees of mutual predictability in our interactions. This way we avoid resorting to requirements stronger than the bounded rationality assumption. He thus appeals to different kinds of mechanisms of 'individual and social construction of higher degrees of mutual predictability'.[35]

He dismisses, in the first place, the 'endogenous solution' (self-enforcing agreements) on the grounds that more unrealistic assumptions would need to be incorporated for it to obtain, like low-cost choice among partners, low-cost monitoring of the past performance of potential 'partners', and no information costs associated with idiosyncratic perceptions and preferences.

---

[35] Scharpf, 1990.

These latter, for example, draw too much on the 'communicative capacities of real life actors who would need to recollect, specify, transmit, and correctly interpret all decision premises that are potentially relevant in characteristically ambiguous interaction situations.'[36]

So, he argues, we need institutions, rules and conventions in order to reduce the information costs as well as the vast range of feasible strategy options. These mechanisms substitute 'common sense' for common knowledge. They make cooperation possible but not necessarily efficient.

Still, choices are constrained by a subjective element, namely, 'actors' perceptions of the situation and their preferences among outcomes.' Scharpf undertakes to differentiate this motivation from self-interest, insofar as this latter may be externally constrained and 'internally shaped by socially constructed criteria of relevance and rightness', like norms of universalistic justice, morality and fairness, 'or in terms of socially defined roles and collective identities to which actors must refer for their self-interested evaluation of outcomes.'[37] In addition, another mechanism would be the networks or 'relation-specific orientations', which may have a positive effect on the predictability of the overall interaction.

Scharpf acknowledges that the consideration of these mechanisms greatly enhances the complexity of our theories, yet he is not willing to deduce them 'endogenously'. It is something we should give up, he suggests.

Another direction is given by the evolutionary approach to games, which has developed from the seminal writings of Maynard Smith (1982) and Axelrod (1984). According to Kirman (1992), this approach is seemingly more convincing than classic non-cooperative game theory in that it remains faithful to GT's original 'radicalism'. In fact, in its early years, as we have seen, GT proposed to deal with a different phenomenon from the one conventional economic theory had so far dealt with, namely the teasing question of how **heterogenous** individuals might coordinate their actions so as to give rise to an overall order. The image of a representative individual, as in conventional economic theory, as well as its equivalent in the classic non-cooperative game

---

[36] Scharpf, 1990, p.482/483.

[37] Scharpf, 1990, p.485.

theory, which makes different individuals collapse into the image of a well-informed and 'commonly known to be rational' agent, miss the point of coordination: there is no question of coordination when you just have one single agent. A less conventional and yet game theoretical view of the economic system is thus expressed:

> In reality, individuals operate in very small subsets of the economy and interact with those with whom they have dealings. It may well be that out of this local but interacting activity emerges some sort of self organization which provides regularity at the macroeconomic level. (Kirman,1992:132)

Here it is easy to read between the lines a major uneasiness with reductionist efforts taken by standard theory to the effect that any coordination might be inferred from the strategic rationality of its parts. Full-blown rationality is to be replaced with less perfect forms, such as imitation, and further stimulated by external forces, like selection. Thus, maximizing behavior as well as common knowledge are dismissed. Economic agents strive to adapt to their environment through a process of trial and error, based on imitation of the behavior of those who have thus far succeeded. Some evolutionary stable strategies will emerge over time, and chance will select one among them.

Another interesting suggestion is that entailed in Dupuy's (s.d.) notion of 'temps du project' (plan's time) and the kind of rationality associated with it, namely, 'libre arbitre' (free will). It displays a kind of rationality that appears to be 'irrational' according to our instrumental view of it, for it involves the subject's capacity of modifying his past as well as detaching himself from his circumstances. Counterfactual reasoning, self-criticism and self-improvement are (rational) human traits that may and indeed should be accommodated within a broader vision of rationality, as free will, he suggests. Incidentally, Dupuy claims that this view of rationality might help us in understanding the emergence of cooperation in non-cooperative games.

6.Concluding Section

In this chapter I have tried to go through the game theoretical view of social interaction in terms of strategic interaction as well as to pick up some ways in which this strategic interaction was resolved in terms of strategic rationality alone. To this end, I looked into two slightly different approaches to strategic interaction, those of vNM and Moore-Anderson, with a view to finding

out what is it peculiar about a **strategic** interaction. The two approaches appeared to sustain that the behaviourist view of social interaction is ill-founded as it overlooks the fundamental fact that in social life action affects and is affected by possible reactions in a quite unpredictable manner; therefore people always need some 'theory' to go about their affairs, to help them thinking about the ways they may affect and be affected. Where these approaches disagree is over the possibility that these 'theories' converge to a fully articulable 'Theory'. VNM believed that there could be such a theory, determinate enough to provide unique rational ways of going about interaction problems, whereas Moore-Anderson deny this possibility and think of the 'Theory' as just a set of heuristic devices. My guess is that Moore-Anderson have the better insight for that matter, as the question of formation and revision of beliefs in interactive contexts seems to me to be fairly tough and requires one to step into the examination of complex mechanisms of generating mutual predictability.

I then attempted to discuss some of the views of strategic rationality. My interest was not so much to describe equilibrium notions or solution concepts as to examine the question of how wise should people be assumed to be in order to be able to find their way to the equilibrium strategy profile. My attention was then directed to the common knowledge assumption. This assumption is very difficult to make sense of outside the formal apparatus, specially in the way it has been worked out in the context of modern GT. Nevertheless, Lewis has cast some light on it, and Dupuy has investigated some of the intuitions behind it. In the context of the Nash concept, complete information, rationality, and common knowledge are deemed to be sufficient for an equilibrium to arise. Apart from the meagre relevance of the standard Nash solution to solve many strategic problems, it also revealed its overly strong belief-structure which assumes certainty as to the other's optimal choice. The relaxation of Nash's belief structure by Harsanyi has introduced the Bayesian framework into game-theory, admittedly with a view to turning a game-theoretic problem into a decision-theoretic one, as Harsanyi says. Aumann even more radically assumes that uncertainty is widespread to the point of almost eliminating any idea of a 'personal' choice. To cope with uncertainty under this strong sense, people are assumed to be able to calculate subjective probabilities which are to reflect their beliefs concerning the behavior of others. In order to ensure that different people will think their way to the equilibrium profile, Aumann introduces the Bayes's rule and the assumption of common priors. In this way, subjective

probability distributions held by different people will converge to a probability distribution held by an outside observer. This is said to impersonate the ignorance the model attempts to deal with. However it appears to be a meta-player displaying a view of ignorance as something quite external to the efforts undertaken by the players to acquire some knowledge of their situation. The question then arises how this position can be justified in terms of the very game-theoretical original objections to an outside perspective?

Consistent maximizing action in the face of uncertainty is certainly not the view espoused by the alternatives mentioned in section 5. Sharpf and Kirman call attention to the social and historical nature of rationality. In these views, much of the information needed for successful strategic interaction is socially provided either by mechanisms intentionally constructed to this end, as in Sharpf, or by evolving institutions that have been somehow 'rationalized' (selected), socially and over time as in Kirman. Scharpf, echoing Moore-Anderson's concerns, would not recommend a fully-embracing theory, able to deduce those mechanisms, as he sides with a view of a more descriptively dense, or complex **theory**, whereas Kirman would like to see a theory of a more complex **object**, his heterogenous agents. A non-instrumental view of reason, or a more complex idea of **reason**, is recalled by Dupuy. All these contributions may be integrated into game-theory in a way that may render it richer. I would however like to address the epistemic problems raised by strategic thinking in terms of a more general **philosophical** outlook.

The purpose of this chapter, then, was not to provide a review of the 'state of the arts' in game theory, neither in terms of its accomplishments nor failures, but to **interpret** GT as producing a particular view of social interaction. My claim is that GT can be interpreted as providing an epistemic-like vision of that interaction, as it emphasizes the extent to which individual action is often enough a reaction to potential action by others and can only be understood in the light of the beliefs, conjectures and expectations held by the acting people with respect to one another. However original and thought-provoking this vision may be, the way the theory has been evolving has turned out to be quite frustrating, since this epistemic vision, short of good philosophy, ends up as a poor exercise in behaviorism.

To think of different people as having to choose among different strategies corresponding to what they expect the other's choice to be suggests an idea of social interaction as an intercourse between at least two distinct individuals, in the sense that one important constraint on someone's choice and its outcome is precisely his awareness that the other's choice will affect him. The exercise of outguessing the other's choice may be made trivially easy to undertake, as one conjectures that there is only one rational way of going about a certain problem and that all that is needed in that search is to describe the decision problem properly and find the equilibrium pair of strategies. However, as this may happen quite rarely, the exercise of outguessing may become a very intriguing one and might even gain some philosophical overtones. In the route followed in this chapter I have called attention to the fact that the strategic interaction as an interaction that involves a great deal of uncertainty, and uncertainty of a very peculiar kind, has prompted a response from the side of Bayesians like Aumann who try to address the crucial issue of how beliefs of different people may come to be aligned and support a suitable equilibrium. Two tools were then presented in favor of the possibility both of an alignment of beliefs and an equilibrium point, namely, Bayes's rule and the assumption of common priors, the use of which I believe still deserves a more solid philosophical justification. It is unclear to me why we should reduce our reasoning process to an inferential mechanism such as Bayes's rule. It is also obscure to me why different people with the same information should hold the same beliefs; perhaps they are not so different. In this case, we may well be in the universe of the 'one' instead. That is why I proceed to examine, in Part III, a different framework, in search for a better philosophically situated position about the importance of knowledge in the social world as well as what we can possibly know.

The suggestion I shall pursue further in chapter 7 is that we should turn to inquire as to how we have come to acquire the knowledge we need in order to act and interact with others, or in other words, what our epistemic condition is. In short, is social interaction a matter of outguessing, bluffing and signalling? My contention is that it is more than this, and that to understand social interaction we should first gain an understanding of the social content of our rationality. So, in acting we resort to condensed forms of knowledge, like rules, institutions, habits and traditions of which we are not (and cannot be) fully aware. These forms constitute both the measure of our ignorance and its redemption. In any case, this perspective may display the advantage of not

taking the 'knowledge-ignorance game' as naively as a zero-sum one, or even a non-cooperative game, in which 'the more I know the less ignorant I am', for it also encompasses the possibility that 'the more I know the less I know', as when we take knowledge as constitutive of our world so that every time one consumes some of it one further produces it.

PART II

AUTONOMY- BEYOND SELF-INTEREST (THE ONE)

INTRODUCTION

In this Part, I take up arguments that propose that the social order is the result of rational (parametric-public) choices made by individuals who are not exclusively motivated by self-interested considerations, in however enlarged a sense. These approaches assume that there can be a motivational sphere that goes beyond self-interest and this they call a **moral** motivation. They also postulate this motivation to be the decisive foundation for a well-grounded social order. The moral motivation is supposed to provide an **objective** point of reference that transcends the individual subjective sphere and objectively connects the individual with his whole community, and mankind. In terms of their main elements, these arguments conceive of the social order as a well-established social unity, underscored by well-grounded interpersonal comparisons where the reasoning needed for such comparisons is provided entirely by the **intra**personal comparisons of **one** rational individual, endowed with a special **moral** inclination.

The label 'autonomy' is meant to indicate the assumption of the moral priority of the individuals that these images of social order proclaim and also, more importantly, that the society is conceived as well-ordered whenever it is so designed as to satisfy **individual ends**, as these are **properly described** in terms of possessing some **moral** properties. In this sense, autonomy means not only that each individual has well-defined **personal** (egoistic or altruistic) preferences over a given set of social alternatives, but also that the preferences over these states may on occasion be **moral** (non-personal) ones, indeed they should be so if we are talking about a well-grounded social order. In a sense, this Part may be seen, looking backwards, as an inquiry into the Kantian alternative view of individual reason, keeping however a forward-looking interest in its feasibility, both from a utilitarian and a contractarian perspective. In relation to chapter 1, this Part aims to go beyond self-interest into what was at the time the mysterious Lockean imperative of making the General Good part of the individual's ends.

Thus, in relation to Part I, this Part addresses the problem of producing a sound argument for non-exclusively self-interested behavior, a question that particularly hampered the account of

political order in chapter 2. Recall that cooperation appeared possible only so long as cooperative behavior was assumed, and this could not be proved to be the unique maximizing behavior. This part takes up the hypothesis that cooperative behavior may have a source other than maximizing considerations alone.

It is one conclusion here that the approach to interpersonal comparisons wholly in terms of the intrapersonal deliberation of one individual considering what the best possible social arrangement is, our world of the 'one', draws heavily on a "rationalistic conception of rationality", to put it in the words of Bernard Williams, and, in this sense, displays a sort of 'abuse' of reason. This means, in our context here, that the **moral** reasoning, the search for an objective viewpoint, was somehow settled in terms of a **theoretical** reasoning. As this route is open to *diaphonia*, indicating plural and conflicting moral principles to rule the basis of our social arrangements, we lack the expected unity here. As a consequence, a different perspective is pursued in Part III. In that Part, the individual deliberation *lato sensu* is assumed not to be a mere matter of method and information, because a genuine question arises concerning our rather limited knowledge capacities. A different link between knowledge and normative purposes is then suggested departing from the reductionist view taken in the present Part.

The exposition of the ideas in this Part pursues the following scheme. Chapter 4 tackles the main characteristics of the second 'type' proposed in matrix 1, the 'parametric-public' model, which sketches the elements of the broader 'autonomy' perspective. This chapter digresses around Kenneth Arrow's social choice theory, as this invites a reflection on a certain way of seeing the social coordination. Explicitly, the insight, borrowed from Arrow, is that if individual ends, to which the social order should be responsive, are interpreted in terms of bare individual or personal preferences over a set of alternative social worlds, where preferences are only constrained in a formal sense, then it proves impossible to attain the social coordination that would result from putting together these preferences in a noncoercive way. I take this as a rather 'inductive' question and the impossibility result as an expression of a kind of 'inductive fallacy'.

Next, in chapter 5, a sort of 'deductive' rejoinder to the above fallacy is presented, stemming from the rule-utilitarian approach by John Harsanyi. Restrictions on the individual ends are proposed in order to overcome the Arrowvian impossibility and reasons for these restrictions are also offered. Two main sets of restrictions deserve mention here, namely, the introduction of a new apparatus for defining rationality (the so-called Bayesian rationality), and the assumption of a moral faculty of human beings, sympathy. Both capacities, it is then implied, are better developed in better educated people, as they are better trained in logical inference, are not supposedly hindered by emotions, and have access to the best information available. In this way, grounds for objective interpersonal comparisons are offered which are based on morally constrained intrapersonal deliberation. The social order is solved as a peculiar rational choice of **one** individual.

Chapter 6 surveys the Rawlsian solution to the Arrow impossibility, as we have interpreted this latter in chapter 4. It is also displayed as a deductive-like approach to that impossibility as it presents arguments for restriction on individual ends drawing both on traditional rational choice view and on a Kantian view of individuals as 'standpoints' endowed with a moral motivation. Two books by Rawls are referred to, A Theory of Justice and Political Liberalism, as well as a number of writings that came to light between their publication. The earlier Rawlsian moral argument is rather a premise, whose backdrop is a certain moral conception of human beings as moral capacities with certain finalities. The 'decisive' argument, issued from a rational choice perspective, is, nonetheless, constrained by the assumed moral backdrop. More sensitive and responsive to the diversity and complexity of individuals' ends, the later Rawlsian argument relies on a conception of public reason, where an agreement on rules to regulate the social coordination is nevertheless to an important extent coextensive with an agreement on theories about social life as well as on procedures and methods of science and common sense (since we would want to have a common evaluation of our institutions as well as to monitor the correct application of the principles, which we all understand in the same way). As in the case of Harsanyi's rule-utilitarianism, a basis for objective interpersonal comparisons is provided stemming from morally constrained intrapersonal deliberation. Similarly, the social order is viewed as a morally constrained rational choice.

A final word is required to state a number of **disclaimers** that call attention to the obvious limits of the task undertaken here. In particular, the discussion around the Arrowvian impossibility is not a technical one, nor is that around the utilitarian theorem in the chapter on Harsanyi. As far as Rawls is concerned, my argument is not a thorough interpretation of the Rawlsian theory (or theo*ries*) of justice, nor a contribution to the debate on the earlier and the later Rawls. I have attempted to make sense of the arguments by these authors as far as my particular question was concerned: I wanted to study their views of social order, from (what I presumed to be) the shared premise of theirs that the social order is a kind of arrangement of individuals' actions towards the furtherance of their ends. To this end, I have concentrated on the kind of **idealization** they assume (how well ordered the coordination of individuals' ends can be) as well as the extent to which this involves constraints on individual ends that have a basis on certain **knowledge** assumptions. In particular, I have attempted to characterize their views in terms of a common pattern, that of taking the social order as grounded in objective interpersonal comparisons (IC) and these as being given to our knowledge through a peculiar deliberation procedure undertaken by any rational individual. In critically assessing this pattern, I indicate the precariousness of making **objective** IC rely on **theoretical** knowledge.

Chapter 4

SOCIAL ORDER AND JUSTICE

1. Introduction

The purpose of this chapter is to set up a provisional frame for a certain view of social order. It draws essentially on Arrow's social choice theory as this yields, in my interpretation, the main elements for that particular view. Thus, this chapter looks into what I call the 'aggregative' or 'parametric-public' model, cell II according to the taxonomy developed in the Introductory Chapter. It will represent an initial frame for the quite 'rationalistic' conception of the social order taken up in the present Part of the dissertation. This Part intends to account for the social order in terms of individual (parametric) rational choices among alternative social states.

A technical discussion of Arrow's theory is skipped, as stress is put on other less developed aspects of it. In particular, the Arrowvian impossibility (a consistent social ordering cannot result from consistent individual orderings of preferences) is elaborated here as a suggestion for the further quest for the difference between the **social order** and a bare **social ordering** of individuals preferences. Arrow was generous enough to append to his impossibility theorem the hope for a way out, in his suggestion that **interpersonal comparisons** should be permitted on the way to a solution; it is only that they need to be justified. This suggestion by Arrow is here worked out while the enterprise of justification of these comparisons undertaken by other authors is examined in chapters 3 and 4.

More to the point, the hypothesis taken up here is that the Arrow result may in some meaningful way be interpreted as suggesting that to justifiably aggregate individuals' preferences one would need a moral argument that properly addresses the issue of interpersonal comparability. The possibility that this aggregation may successfully be achieved through the relaxation of the Arrowvian conditions will not be pursued here.

In section 2, the 'parametric-public' approach is renamed as 'order as justice' to spell out the specific connection between order and justice that is present in an aggregative account of the social order. To this end, some of the premises of Arrow's social choice theorem are outlined (in subsection 2.1) that characterize the broad approach, such as individual **preferences** as the sources of a social state, the kind of justice-wanting conflict that is likely to emerge (distributive issues and conflict of rights), and the kind of solution most conducive to internalizing these undesired 'interactional frictions'. Following that, this section sets up both the possibility of a normative reading of Arrow's impossibility result different from the usual technical one (in subsection 2.2), and a conjecture about a possible conceptual connection between order and justice to which the Arrow theorem gives rise (in subsection 2.3). The Arrow result, it is then argued, evokes the thought that the social order is, at some deeper level, conditioned by a criterion of justice, of the just interpersonal comparisons among its members. In subsection 2.4, the Arrow result is presented and interpreted in terms of the issue of interpersonal comparisons. Of course, in their personal preferences people make their subjective interpersonal comparisons. The question is, How can these comparisons be made and incorporated in the social preferences in a way that respects individuals' autonomy?

Finally, in section 3, an introduction to the issue of interpersonal comparisons from an ethical viewpoint is attempted.

2. Order as Justice

In the following subsection, we will be going through some of the more visible aspects concerning a view of social order as portrayed in cell II, the 'aggregative' or 'parametric-public' outlook. In the subsections that follow, increasingly risky interpretations of that approach are undertaken with a view to presenting some elements of a common paradigm, namely, the 'order as justice' view of social order.

2.1 (Imperfect) Parametric Public Choices

To begin with, the notion of parametric public choices needs some clarification. A 'parametric' rational choice requires the exclusion of a number of aspects from the framing of the choice

situation, as in Arrow's (1951, 1963) proposal: game-like aspects, interaction between preferences and methods of decision, and any irrationality of the individuals in the form of violation of the two requirements of completeness and transitivity of the orderings (whenever more than two alternatives are considered) are to be disregarded.

As for the game aspects, Arrow states that 'no consideration is given to the enjoyment of the decision process as a form of play'.[1] More importantly, he also stipulates that no consideration be given to the game-like phenomenon that people may happen to misrepresent their preferences by their actions.

As for the interaction between preferences and methods, he further states that:

> If individual values can themselves be affected by the method of social choice, it becomes much more difficult to learn what is meant by one method's being preferable to another.(Arrow,1951,1963:8)

Rationality is understood, as he puts it, as maximization of some sort. Finally, summing up his initial qualifications, he puts them together in the following assertion:

> In addition to ignoring game aspects of the problem of choice, we will also assume in the present study that individual values are taken as data and are not capable of being altered by the nature of the decision process itself... Finally, it is assumed that all individuals in the society are rational.(Arrow,1963:7-8)

For our purposes, this shall suffice to identify 'parametric' rationality in Arrow's decision-makers.

On reflection, it seems that the above exclusions aim to reinforce the autonomy or the 'decisiveness' of **individual** decision-making, as well as its coherence, in terms of connecting individual choices to the preferences of **rational individuals** alone (to repeat, to the exclusion of strategic thinking, interactions between ends and means, and irrationality).[2]

---

[1] Cf, Arrow, op.cit., p.7.

[2] A related question is whether 'preferences' can be the substrate of such autonomy, as Elster (1982) has pointed out (Sen,1970,1985a; Broome,1991), because of cognitive cacophony. For example, people may happen to adjust their preferences relying on considerations of feasibility. Sen (1970, 1985a), in turn, argues that preferences should not be

(continued...)

The 'public' nature of the issues that are the object of the preferences of the individuals, in turn, is indicated by the fact that they configure a set of alternative 'social states', thus defined by Arrow:

> a social state is a complete description of the amount of each type of commodity in the hands of each individual, the amount of labor to be supplied by each individual, the amount of each productive resource invested in each type of productive activity, and the amounts of various types of collective activity, such as municipal services, diplomacy and its continuation by other means, and the erection of statues to famous men.(ibidem:17)

So, Arrow argues that the preferences over social states are still individuals' but their content does not of necessity reflect egoistic inclinations but may well accommodate values to which the individuals happen to subscribe, other than mere personal welfare considerations. Put somewhat differently, preferences are autonomous (in the sense that they are generated in the individuals), but their motivation is not necessarily so (in the sense that the preferences also express aims other than the individuals' direct advantage).

Indeed, Arrow is clear enough about this point and insists that 'tastes' or 'egoistic desires' cannot be relied upon as guidance for choices among a certain class of alternatives, for example, the amount of public expenditure, or its distribution among different items that have no direct bearing on the individual's welfare, or some particular pattern of equity the individual may happen to desire of the alternatives. In these cases it is natural to assume, Arrow argues, that individuals refer to their values, not only to their tastes.

However, Arrow also indicates that different tastes and values are sources of conflict when many wills are involved. Possible objects of conflict are different profiles of income distribution, he admits, and 'other-regarding' preferences (the preferences one has regarding the behavior of others), a point stressed by Sen (1970).

(...continued)
viewed as such a substrate, because of rights and distributive issues.

There is a natural way of overcoming such seemingly intractable differences, thus waiving these 'interactional residues' (distributive and rights issues), which is to suppose the existence of a 'general will', incidentally an image that abounds in political philosophy. As Arrow recalls:

> The assumption ... of complete agreement among individuals on the ordering of social alternatives, may seem obviously contrary to fact. But, properly interpreted, it is at the basis of a great portion of political philosophy, namely, the idealist school. The fundamental doctrine of the group is that we must distinguish between the individual will, as it exists at any given instant under varying external influences, and the general will, which is supposed to inhere in all and which is the same in all; social morality is based on the latter. This view is expressed in the works of Rousseau, Kant, and T.H. Green, among many others.(ibidem:81)
>
> The idealist doctrine then may be summed up by saying that each individual has two orderings, one which governs him in his everyday actions and one which would be relevant under some ideal conditions and which is in some sense truer than the first ordering. (ibidem:82/83)

Moreover, Arrow asserts that some sort of consensus on the ends of society is indeed necessary 'or no social welfare function can be formed', and it is to be found either in the individual orderings or 'in the moral imperatives of various members of society'.

What could possibly be the reasoning in support of the idea of the splitting of the will into two different ones, in order to find the 'ends of society'? Rawls and Harsanyi address this question: they frame the social choice, the choice of a social state, as to somehow force the individual into choosing, abstracted from his circumstances but still taking his own advantage into account and in use of his reason, an ideal society, one in which he, whoever he happens to be, would be willing to live.

Put this way, the ideal society can no longer be fully specified, as were the Arrowvian social states; in fact, this is not the expected result of individual decision making. What is actually expected as an outcome of this kind of mental process is a principle or a set of principles that should **order** such social state.

It may be argued that, different as they may appear to be (and really are, on a number of accounts that are disregarded here), Arrow and Rawls-Harsanyi are only departing from diametric points

along the same line of reasoning: the former, from actual (or weakly constrained) individual choices to desirable social states; the latter, from individual choices under ideal conditions to desirable social states. Arrow wonders whether rational people unconstrained in any sense other than their rationality may not shape a rational society in a somewhat axiomatic way, and he concludes with an impossibility result, which he suspects may well be an expression of what are here called the interactional frictions or residues. Rawls-Harsanyi wonder what the conditions are for people to succeed in such a construction; theirs is the standpoint of an ethical observer that will supply the necessary conditions (and in this sense modify the initial set of conditions imagined by Arrow) for people to design an ideal order. Indeed, it is possible to characterize Arrow's parametric-public model in terms of an **imperfect** parametric-public choice, to contrast it with the more developed forms worked out by Rawls and Harsanyi where there are almost no interactional residues left and the social state is designed in a more abstract way.

2.2. Normative Social Choice

In his Social Choice and Individual Values, Arrow concludes that the passage from the individual level to the level of society is not rationality-preserving, no matter what (collective) procedure is adopted. At least when rationality is defined as a weak-ordering (where suitable axioms apply, viz., completeness and transitivity), the autonomy of the choices is assumed (which is to be guaranteed by a set of conditions), and interpersonal comparisons (IC) are excluded.

An immediate reading of the theorem is that a set of reasonable conditions to be imposed on any (collective) procedure that is supposed to link individuals' preferences to social states will **conflict** over an unrestricted domain of the individuals' choices. Therefore, unrestricted (although well-behaved) preferences of the individuals, when taken together, could lead to collective chaos.

Still, one may ask, what is this sort of reversed composition fallacy a symptom of? If we discard the case of improper description of the social coordination,[3] two alternatives have been taken up,

---

[3] This alternative is pursued in Parts I and III of this dissertation. In Part I, for instance, the social coordination is described in terms of strategic interaction, and in Part III in terms of a non-reductionist 'interacting individuals' interpretation.

respectively by social choice theory and moral theories, namely, that the resulting irrationality is a symptom either of lack of logical depth on the part of the theory or of ethical shallowness.[4]

Accordingly, one way out is to look for 'possibility' techniques, in the form of second thoughts regarding the initial Arrowvian 'reasonable' constraints on the procedure. Consequently, relaxation of these conditions has been widely pursued in the literature, motivated by the view that one should restore the **consistency** conditions of the social utility function.

An admittedly less natural way out is to skip the 'impossibility' issue in its logical aspects and look at a practical form of the problem, as for example a conflict among a certain set of **values**, where the conditions are interpreted in somewhat normative terms, that demands that we take a different step than possibility techniques in order to **decide** on a solution. In Hylland's (1986) suggestion, we would then be looking for the standpoint of an ethical observer who is properly to **judge** the values that should be kept and the ones that could be discarded. Again, the idea here is that the formal conditions should then be interpreted in normative terms.

In this particular way, being provoked by the impossibility result one might be required to go deep into ethical considerations. In particular, it might be of interest to identify and come to grips with the more likely sources of conflict, as the literature has done, for example the distribution of income or rights issues (Sen 1985; 1970).

Indeed, Arrow himself ventures on a direct connection between the social choice theory and a theory of justice, in his Social Choice and Justice (Essay 11). Though he indicates that the domains of each of these theories are clearly not the **same**, he nonetheless believes in an **intimate** relation between them. So, on the one hand, he says, if one takes the Pareto criterion as part of the definition of a satisfactory theory of social choice (Arrow,1984:147) one already has a 'distributive' constraint on the feasible set of alternative social states. On the other hand, a theory

---

[4] Hylland (1986), for instance, accepts the following view: "On one level, social choice theory is a mathematical discipline; conditions are formulated precisely in the form of axioms, and theorems proved. But if the theory is to be of any use, the story cannot end there. The results must tell us something about 'real' issues." (Hylland,1986:45)

of justice can be said to result from a collective decision (even if this decision favors some decentralized mechanism of distribution).

2.3. Aggregating and Weighing: order as justice

I want to take up and build on Arrow's suggestion of an intimate relation between the social choice and the justice theories. My own view is that axiomatic social choice gives an insight into a peculiar **conceptual** connection between order and justice, intimating an even closer relationship than that spelled out by Arrow between the social choice theory and the domains of moral and political philosophy.

It may be argued that the conceptual connection between order and justice is twofold. On the one hand, Hardin (1988) argues that justice can be understood as order. He stresses the difference between what he calls 'justice as order', where justice is seen as a synonym for order, and distributive or social justice; it seems, he still argues, that the notion of justice has now acquired a chiefly distributive connotation but it has also been

> and ... very clearly a notion of order. The thirty-first Principle Doctrine or Sovran Maxim of Epicurus is: 'The justice which arises from nature is a pledge of mutual advantage to restrain men from harming one another and save them from being harmed' (Whitney J. Oates, ed., The Stoic and Epicurean Philosophers, New York: Modern Library, 1940, p. 37, cited in Hardin, 1988:36fn).

On the other hand, what seems to be characteristic of the approaches in cell II is precisely the conceptual autonomy and **priority** of justice over order, in the sense that, ideally, the social unity is conceived as springing solely from a prior criterion of justice in a broad sense, that is, as a rule of adjudication of conflicts concerning the distribution not only of rights and duties as in a 'justice as order' view, but also of economic and social advantages. The resulting principles of justice are the outcomes of a social choice, and are to work as constraints to be imposed on the set of social states rather than thorough social states. The theories located in cell II are interested in articulating this criterion; they also want to evaluate a conception of justice with respect to some desirable attributes it should have. These we will call theories of 'order as justice', just to emphasize the above priority.

I intend here to capture the peculiar connection between order and justice presented by approaches in cell II through the idea of 'aggregation'. In this way, images of social order as the result of choices performed by parametric-rational individuals share the assumption that the social order is the outcome of an 'aggregation' of largely autonomous individuals, where strategic and interactional considerations of the means-ends kind are excluded. In particular, the idea of the social order as an aggregation here presupposes the existence of a prior conception of justice in the broad sense.

Aggregation is quite straightforward as far as Arrow's argument is concerned, for he utterly commits himself to the search for a rationality-preserving **aggregation** procedure of individual preferences, like voting, to design a social state.

As far as Harsanyi and Rawls are concerned, the idea of aggregation still arguably underlies the image of a well ordered or just society, as this is regarded as a well arranged gathering of individuals which are seen as self-generating sources of aims and claims. In particular, society is seen as the assembling of individuals who are expected to somehow use their equal capacities as rational beings to accommodate their different concerns. As a consequence, and making the most of these 'equal capacities', the aggregation procedure that takes place, replaces the set of **many** and diverse individuals with just **one** individual, through specific mechanisms, namely, the Rawlsian device of the original position, whereby an individual's reasoning process (when properly constrained) comes up with the two principles of justice, and the Harsanyian ideal observer, where the preferences of a sympathetic spectator lead to the utilitarian principle.

We may wonder, however, what the path is to achieve that 'replacement'. A crucial question relates to the 'brute fact' of the heterogeneity of individuals's ends. In constructing the aggregation procedures, these approaches have to address the question of how to find a common ground in order to accommodate the differences between the individuals, so as to reach a coherent picture of society. The implication is that social unity involves more than the sum total of individuals' ends, on the basis of some minimum common denominator (let us say, that everyone wants to seize the benefits of social cooperation) as these ends may conflict; it also involves a preceding and articulable common notion of how to solve the likely conflicts that are to emerge on account

of the differences between them. 'Order as justice' gives primacy to this latter aspect. According to this approach, social order is dependent upon a prior proper weighing of different people's aims and claims, prior principles of justice, that is. To aggregate we need to weigh. To justify our aggregation we need to know and be able to justify the criterion for interpersonal weighing that we have used.

The hypothesis worked out here may be restated as follows: the 'order as justice' approach elaborates the idea that a conception of social order as an aggregate of many and diverse individuals aims, however interpreted, presupposes a criterion of justice (as implied in the idea of a distribution of weights); it should also be able to develop such a conception so as to fully articulate the criterion in terms of some general principles. This criterion of justice itself results from a peculiar social choice. In other words, to achieve the social preferences we would need to have established a criterion of justice, of the just interpersonal comparisons of people's preferences, in the first place. This criterion itself can only result from a specific social choice where considerations other than personal preferences, be they egoistic or altruistic, apply. This aspect is worked out by Harsanyi and Rawls.[5]

---

[5] This characterization may encompass to a variable extent many foundational works in social choice theory (Arrow and Sen's, for example) as well as the theories of justice of John Rawls and John Harsanyi, however different their conclusions happen to be. In any case, the largely inductive way whereby social choice theory proceeds on the way to the aggregate format of social states contrasts with the largely deductive manner adopted either in the Rawlsian construction of a reasonable viewpoint or in Harsanyi's moral standpoint. These latter perspectives are to provide restrictions as well as reasons for restrictions to be imposed. (However, one must not exaggerate the contrast, for social choice efforts to derive social states from individuals' preferences restrict our attention to preferences only, and also already embody a (weak) restriction on these preferences in the form of the rationality axioms already mentioned. In this sense they are somewhat deductive from the onset; whereas, Rawls and Harsanyi's views might be regarded as somehow contaminated by an unconfessed piece of inductivism, and of intuitionism it might be argued, as particular beliefs entertained by the authors are elevated to the status of shared beliefs, although not always in a cogent way.) So, it is of interest here the fact that the 'inductive' procedure ends up with an impossibility, an induction fallacy of sorts, escape from which seems to be pursued by 'deductive' approaches. How well these approaches have managed to overcome the problems of the former without ending up with a renewed impossibility of their own is open to debate. (In fact, I contend in later chapters that a renewed induction fallacy is likely to arise in the search for objectivity undertaken both by rule-utilitarian and Rawlsian approaches.) Another way of interpreting the connection between the Arrow result and deductive approaches, in a more integrative key, is to regard the former as clarifying what the issues are that need further (deductive) argument for an individualistic approach of the kind considered here (Parametric-public) to succeed in giving rise to a cogent image of social order. It is worth adding, however, the following qualification: as far as Arrow's theorem is concerned the assumption of individualism is understood in the sense that social choices are seen exclusively as the result of individual (rational) choices and these, in turn, are to stem only from individual preferences over a set of given alternatives. The deductive arguments to be examined shall draw either on a preference-scheme, as in Harsanyi's approach, or on a non-preference scheme (at least not in terms of 'utility', such as Rawls'.

Let us now turn to Arrow's teasing problem in order to pave the way for Rawls and Harsanyi's contribution, as they will appear in the next two chapters.

2.4. Social Ordering and Social Order: Arrow's problem and the IC issue

In connection with the preceding section, the purpose of this subsection is to take on an interpretation of the Arrow problem in terms of the issue of interpersonal comparisons. This is not to deny that the Arrow impossibility is the formulation of a logical inconsistency that follows a certain description of how social preferences are formed, as stressed by Samuelson (1967). However, Arrow's statements on different occasions reinforce the impression that his ultimate aim while working on his social choice theory was mainly normative. In particular, as he declared (Arrow,1984:160), he was interested in providing a basis for a theory of justice from a social choice theory framework. It is to this particular aspect of his impossibility result that we now turn.

* * *

Arrow's General Possibility Theorem states that there is no collective procedure that connects individuals' preferences over social states to a particular social state, in a satisfactory way. By collective procedures he means those decision methods that are characteristically unimposed (by external forces like customary rules) and non-dictatorial (someone's preferences prevailing, regardless of what the others' preferences are), such as markets and voting; although he only concentrates on voting procedures. By a satisfactory way he means a rational way implying that the procedure should abide by the requirements of completeness and transitivity of the ordering of the alternatives (a set of social states) and satisfy a number of other conditions (originally five, and, later, in the 1963 revision, four).[6]

---

[6] The above mentioned conditions for the construction of the collective decision device, the social welfare function, are, as reported by Sen, the following: "**Unrestricted domain** (hereafter, condition **U**) demands that the domain of the social welfare function must include all possible individual preference profiles (i.e., no matter what preferences the member of the society holds, the social welfare function can successfully aggregate them into a social preference ordering). The **Pareto Principle** (hereafter, condition **P**) demands that if everyone prefers **x** to any **y**, then **x** is socially preferred to **y**. **Independence of irrelevant alternatives** (hereafter, condition **I**) requires that the social ranking of any two states **x** and **y** depends only on the individual rankings of these two states. (...) Finally, **non-dictatorship** (hereafter, condition **D**) prohibits the presence of a dictator (i.e., a person such that whenever he prefers any **x** to any **y**, the result is that **x** is socially preferred to **y**). Arrow's impossibility theorem states that if there are at least three distinct social states and the set of individuals is finite, then there is no Social Welfare Function satisfying conditions **U**, **P**, **I**, and **D**." (Sen,1985a:1766)

Arrow's **formal** question, to which he devotes his 1951 book, is whether:

> [it is] formally possible to construct a procedure for passing from a set of known individual tastes to a pattern of social decision-making, the procedure in question being required to satisfy certain natural conditions? (Ibidem:2)

Arrow's reportedly **informal** question is: can the consistency achieved at the level of the individual 'be attributed to collective modes of choice, where the **wills of many people are involved**?'(my emphasis).

In a different statement, Arrow also warns that the resulting impossibility comes as a natural consequence of its own starting point, viz. the rejection of Interpersonal Comparisons (IC) of utility. The impossibility is then restated in the following form:

> If we exclude the possibility of interpersonal comparisons of utility, then the only methods of passing from individual tastes to social preferences which will be satisfactory and which will be defined for a wide range of individual orderings are either imposed or dictatorial. (Arrow,1963:59)

Of course, we could rephrase Arrow's above statement in terms of any of the five (or four) conditions for a satisfactory social choice that he stipulates, and select this particular one as 'responsible' for the impossibility result and then undertake its relaxation on the way to a solution, as many authors have done thus far.[7] However, Arrow selected the prohibition of IC in this passage and this seems to be meaningful in our context here. The exclusion of IC is explicitly stated by Arrow as a sort of background constraint that seems to have for him a **different status** from the other axioms. This problem seems to be related to the way he puts his informal question, referring to the difficulty in achieving consistency 'where the wills of many people are involved'. He appears to be interested in knowing the possibility of this achievement without resorting, however, to a premise which he finds very difficult simply to assume, unlike for the other axioms:

> In a way that I cannot articulate well and am none too sure about defending, the autonomy of individuals, an element of mutual incommensurability among people, seems denied by the possibility of interpersonal comparisons. (Arrow,1984:160)

---

[7] I am grateful to Alan Kirman for calling my attention to this particular point.

So, overcoming the impossibility result in the strict Arrow context might require the acceptance of IC, a step Arrow himself was reluctant to take fully.

3. Interpersonal Comparisons: a preview

Arrow declared that economists *qua* economists are not entitled to undertake interpersonal comparisons. Still, he also argued that the exclusion of these comparisons (by the way, the received view in economics and welfare economics at the time Arrow completed his demonstration), and given his idea of a satisfactory method of aggregation, would necessarily lead to the impossibility result. IC are not justifiable, yet without them social choice is an impossibility.

Since Arrow's proof, however, welfare economics has launched a deep change in its stance concerning IC. Actually, the Arrow result provoked a fertile reaction in the form of subsequent work by a great number of authors in the field, building on his suggestion that the IC constraint should be dropped on the way to a solution.[8] (Arrow himself worked out an ordinal approach to IC, the lexicographic maximin or 'leximin' as a social welfare criterion.[9]) As a consequence, interpersonal comparisons have since become an infatuation in social choice theory.[10] (In particular, as they were elaborated on with the tools of Bayesian decision theory, IC were given a ground that could in turn support the utilitarian moral principle, as discussed later in chapter 5)

However, one is entitled to ask, echoing Weber and Hirschman's *pathos* on a different matter: how did it all happen? How is it possible that after decades of deep aversion, interpersonal comparisons have become widely accepted and incorporated, and, finally acquired the status of an almost incontrovertible axiom? The answer, here as elsewhere, lies in the realm of ethics, as

---

[8] Relaxation of some of the other axioms has also been attempted, but I will not focus here on these alternative ways out.

[9] The leximin criterion entails 'judging the social welfare of a state by the well-being of the worst-off individual; if two states tie in this respect, then by the well-being of the second worst-off, and so on'. Cf. Sen, 1985a, p.1772.

[10] Sen reports on the matter as follows: "there has been a great revival in the 1970s in making and using interpersonal comparisons, and for this purpose Arrow's original social choice format has been appropriately adopted. Arrow himself contributed to this revival through some remarks in the second edition of his book (Arrow, 1963, pp.114-15), following an important paper of Patrick Suppes (1957)." (Sen, op.cit., p.1771). Hammond (1991a) provides a detailed reference list on IC of utility.

the main reason for such aversion stems from ethical considerations. IC have been increasingly regarded and portrayed as a crucial foundation for a properly ordered social unity, at least by people persuaded of the 'parametric-public' model, after the usual caveats have been properly addressed.[11]

Closer inspection reveals that the original rejection by Arrow of interpersonal comparisons follows his similar rejection of cardinal measures of utility, be they individual or interpersonal.

> The viewpoint will be taken here that **interpersonal comparisons of utilities** have no meaning and, in fact, that there is no meaning relevant to welfare comparisons in the measurability of **individual utility** ... If we cannot have measurable utility, in this sense, we cannot have interpersonal comparability of utilities *a fortiori*. (ibidem:9, my emphasis)

Many authors before Arrow had questioned the meaning of such measures, the selection of which always involves a decision that is taken beyond the realm of economics. On this view, an economist *qua* economist is not entitled to proffer judgments on welfare policies (as much as measure individual utilities), for the substance of what he is to judge surpasses its **disciplinary**

---

[11] See chapters on Harsanyi and Rawls. Arrow also undertakes an ordinalist IC exercise. On the similarity between Rawls and Arrow's view of IC, I quote Arrow, reporting on the leximin solution to IC of utility: "The work I am reporting on here has an ironic relation to Rawls' difference principle. Under certain epistemological assumptions about individual utilities, a social choice approach leads to Rawls' difference principle - but in terms of utilities, not primarily goods." (Arrow,1984,Essay 11:149) Also Sen's remarks stressing the above similarity are worth quoting: "The theorems presented by Arrow are about getting scalar values for u-vectors (given neutrality, i.e., ignoring non-u features), and the axiomatizations tell us when we should sum, when we must look at the minimum value, and so on. They are not specifically about utility vectors at all." (Sen, 1985, 1773) An important difference between them refers to a somewhat impersonal character of the Rawlsian 'primary goods', in contrast to the rather personal dimension of 'utilities':"It can be asked in this context whether Rawls' indices of primary goods may not be usable in exactly the same way as utilities in the form of u-values in the analysis presented by Arrow. One problem arises from the fact that, strictly interpreted, the holding of primary goods is not a feature of a person's state of existence, but of the means to his achieving one state or another. The contrast between intra-personal and inter-personal comparisons ... cannot strictly speaking arise with the primary goods index. If the primary goods bundle held by A has higher value than that held by B, then we do not know that in this perspective A is simply more advantaged than B is. The contrary case, which can easily arise with advantage interpreted as utility, to wit: UB(xA)>UB(xB)>UA(xA)>UA(xB), simply cannot arise with holdings of primary goods. If xA has a higher index value than xB, then the value I of advantage of the person, interpreted in terms of holding primary goods, will be quite independent of everything other than which bundle the person holds, that is IA(xA)=IB(xA)>IA(xB)=IB(xB). Utility has a 'personal' dimension that indices of primary goods do not; the latter have to be 'impersonal', in this sense."(1773)

competence and tackles essentially ethical issues.[12] This is, in general, the position defended by Lionel Robbins:

> I still think, when I make interpersonal comparisons ... that my judgments are more like judgments of value than judgments of verifiable fact ... I think that the assumption of equality comes from outside [of economic science], and that its justification is more ethical than scientific. (Robbins,1938:660-641, cited in Hammond, 1991a)

This position of Robbins had been pursued by economists for decades before cardinalism's revival in the 1970s.

We can trace Robbins's argument back to its origins in the work of Stanley Jevons (1965). In The Theory of Political Economy, Jevons works out the limits of possibility as far as measurement of economic phenomena is concerned, and these do not seem to accommodate interpersonal comparability, at least of an hedonic kind. Minds are always opaque to one another, Jevons contends:

> there is never, in any single instance [in Jevons's theory], an attempt made to compare the amount of feeling in one mind with that in another. I see no means by which such comparison can be accomplished.
>
> Every mind is ... inscrutable to every other mind, and no common denominator of feeling seems possible. (Jevons,1965:14)

Additionally, aspects of difficult **intrapersonal** measurement were stressed by Jevons:

> Pleasures, in short, are, for the time being, as the mind estimates them; ... It is true that the mind often hesitates in making a choice of great importance: this indicates either varying estimates of the motives, or a feeling of incapacity to grasp the quantities concerned. (ibidem:13)

As a consequence, Jevons argues, it is difficult to assert that 'one pleasure is an exact multiple of another', and this limits his theory to the so-called 'marginal' scale, in the sense that the theory cannot involve

---

[12] Though, of course, after a set of values has been given, the economist can always evaluate their coherence in terms of the formal apparatus with which he is endowed. This point is emphasized by both Arrow and Robbins, op.cit..

> the comparison of quantities of feeling differing much in amount. The theory turns upon those critical points where pleasures are nearly, if not quite, equal. (ibidem:13)

In summary, Jevons's skeptical viewpoint underlines an empirical impossibility as far as a hedonic view of utility is concerned. The implication seems to be that without empirical support no judgment of the kind implied by IC can be uttered at all.

Thus, the acknowledgment of a certain 'empirical opacity' with respect to interpersonal comparisons gave rise to two rather distinct attitudes, represented here by Robbins and Jevons's statements. In the case of Jevons this has implied an abandonment of ethical concerns as hopeless, whereas in the case of Robbins it has issued a recommendation that economists should not meddle in value judgments, and ought, in economics, to remain attached to judgments of verifiable facts alone. This latter seems also to be Arrow's view, which he spells out in the context of his essay on 'interpersonal ordinal comparisons':

> if your satisfaction depends on some inner qualities that I do not possess, then I really have not had the experience which will enable me to judge the satisfaction one would derive from that quality in association with some distribution of goods. Hence my judgment has a probability element in it and therefore will not agree with your judgment. But it is essential to the present construction that the [interpersonal] comparisons (...) be the same (...)(Arrow,1984:160)

Addressing mainly the former contention, as a political and moral philosopher John Rawls states that this kind of skepticism concerning IC is disputable:

> skepticism about interpersonal comparisons is often based on questionable views: for example, that the intensity of pleasure or of the enjoyment which indicates well-being is the intensity of the pure sensation; and that while the intensity of such sensations can be experienced and known by the subject, it is impossible for others to know it or to infer it with reasonable certainty. Both these contentions seem wrong. Indeed, the second is simply part of a skepticism about the existence of other minds, unless it is shown why judgments of well-being present special problems which cannot be overcome. (Rawls,1971:91)

Accordingly, this does not rule out the study of other foundations for these judgments, and the motivation might well be the fact that people make IC daily. Of course, as Rawls has pointed out, these 'daily IC' do not necessarily have a sound basis:

> Simply because we do in fact make what we call interpersonal comparisons of well-being does not mean that we understand the basis of these comparisons or that we should accept them as sound. To settle these matters we need to give an account of these judgments, to set out the criteria that underlie them.(ibidem:90)

John Harsanyi, in turn, rejects the hedonic approach to utility and takes the view that interpersonal comparisons of preferences can meaningfully be made on the way to the social preferences, by an ideal observer, a theoretically constructed standpoint. This device would restore the consistency of the social utility on the basis of sound moral arguments, so he claims. It is to these arguments that we now turn.

Chapter 5

ORDER AS JUSTICE (I): JOHN HARSANYI'S RULE-UTILITARIANISM

1. Introduction

According to the argument developed in the previous chapter, a well accomplished view of social order in rational parametric-public terms would require a full articulation of the question of interpersonal comparisons (IC). That is to say, according to the undertaking in cell II, in order to conceive of the social order as a satisfactory aggregation of individuals' preferences over alternative social states it is necessary to provide a proper view of IC.

Harsanyi's rule-utilitarian view provides one such an articulation. In particular, he produces a justification of the IC entailed in a social state wholly in terms of a specific intrapersonal comparisons process undergone by a rational individual. In other words, towards the construction of a social state from individual preferences people are to be compared in a way they would individually agree, were they to undertake the appropriate hypothetical reasoning process.

In this chapter, Harsanyi's version of the utilitarian principle as a 'principle of justice', of the justifiable IC, is examined. This principle essentially says that in order to construct social preferences people should be assessed equally after their preferences have been established and compared in a proper way. Harsanyi argues that the utilitarian principle is the result of the reasoning process to a principle of justice undergone by any rational individual under appropriate hypothetical conditions. Towards the end of reducing the interpersonal dimension into the intrapersonal, the question of 'order as justice' is reduced into the question of individual rational choice under risky or uncertain circumstances.

This approach, therefore, provides a very reductionist view of the social order in the sense that this latter is understood in terms of individual rationality alone. The social order is, then, built up in terms of an aggregate or **social utility**, which is the weighted sum of individuals' preferences, where the weights are assigned by an ideal or sympathetic observer and the individual preferences

are properly calibrated. Furthermore, society's institutions are to perform so as to maximize that aggregate.

In section 2, the utilitarian principle is presented here under two versions, the veil of ignorance argument and the axiomatic Bayesian argument. Some of the axioms of this latter are discussed in later sections, along the following lines.

The knowledge assumptions regarding the individual utilities as well as the possibility of interpersonal comparisons are inspected in sections 3 and 4, respectively. It is remarked, in section 3, that 'preference autonomy' relies heavily upon a theoretical reason, complementing the instrumental role of rationality with some 'truth' achieved through theoretical means. It is also argued, in section 4, that the proposed IC view rests importantly on the premise that people's thought processes tend to converge to an objective standard in a particular way. Section 5 reacts to the proposed link between uncertainty or risk and moral reasoning in the context of the Harsanyian defense of the utilitarian principle. This section indicates that the two versions of the utilitarian principle presented by Harsanyi are very closely connected.

2. Harsanyi's utilitarian theorem

According to the classical Benthamite utilitarian precept, society should maximize the happiness of its members, or should seek the 'greatest happiness of the greatest number'. Under the 'total utility' approach this is translated into the rule of maximization of social utility, utility meaning a measure of the hedonic substance which society is supposed to maximize. A modern version of the old canon states that society should try to maximize its average utility.

Though the two definitions are mathematically equivalent, they may, however, produce different policy recommendations whenever the population is not held constant.[1] What is more, the updated version of the utilitarian principle takes utility not as a **measure** of the net balance of pleasure and

[1] Cf. Harsanyi (1982), p.46.

pain but as a mere mathematical **representation** of people's preferences, following Pareto and Hicks's contributions to modern utility theory.[2]

As far as modern economics is concerned, the shift from happiness to revealed preferences has essentially meant that the theory is freed from the subjectivist burden involved in its hedonist psychology and introspection. Accordingly, it acquires a more empirical content, for 'people's preferences are revealed by their publicly observable behavior'.[3] People reveal their preferences through their choices. Therefore, a rational economic agent is only required to observe some consistency conditions when mapping his preferences, for his behavior to have economic meaning, and hence to be representable by a utility function.

Harsanyi believes, however, that additional more stringent conditions than the standard revealed preferences approach are required of an individual when it comes to welfare and ethical considerations. We need then to map the 'social preferences', and indeed to give them meaning, starting from the preferences of the individuals. What preferences should we consider? What are the conditions we should impose on them?

According to Harsanyi, ethics is primarily a branch of 'the general theory of human behavior'. It is concerned with rational behavior towards a particular end, as in 'a theory of rational behavior in the service of the common interests of society as a whole'.[4] The implication, he contends, is that as far as ethics is concerned 'the secondary definition of rationality is in terms of maximizing the **average utility** level of all individuals in the society'.[5] Why so? Let us follow Harsanyi's argument towards that end.

In his 1982 article, Harsanyi presents his theory of morality as a branch of the general theory of rational behavior by postulating it in terms of a theory of the rationality of our moral judgments.

---

[2] Cf. Harsanyi (1992), p.1.

[3] Cf. Harsanyi (1992), p.2.

[4] Cf. Harsanyi (1982), p.43.

[5] Cf. Harsanyi (1982), p.44, my emphasis.

Moral judgments, he says, are those judgments which are expected to achieve certain standards of universality and impartiality. In what sense are our moral judgments, thus defined, rational?

First of all, a moral judgment is a judgment of preference. However, it is a preference of a 'very special kind'; it is not expected to reflect the kind of preference we would hold only as rational and self-interested persons. In order to achieve the standard of universality or impartiality involved in a moral judgment, we would need to impose certain constraints on our preferences.

> Even if both an individual's social welfare function and his utility function in a sense express his own individual preferences, they must express preferences of different sorts: the former must express what this individual prefers (or rather, would prefer) on the basis of impersonal social considerations alone, and the latter must express what he actually prefers, whether on the basis of his personal interests, or on any other basis. The former may be called his 'ethical' preferences, the latter his 'subjective' preferences. Only his 'subjective' preferences (which defines his utility function) will express his preferences in the full sense of the word as they actually are, showing an egoistic attitude in the case of an egoist and an altruistic attitude in the case of an altruist. His 'ethical' preferences (which define his social welfare function) will, on the other hand, express what can in only a qualified sense be called his 'preferences': they will, by definition, express what he prefers only in those possibly rare moments when he forces a special impartial and impersonal attitude upon himself. (Harsanyi,1980:14)

Note that this does not mean that we are not choosing **from** a self-interested perspective, as when information about the content of our self-interest is missing we could possibly achieve an impersonal attitude. In fact, Harsanyi claims, a self-interested perspective with the addition of factual ignorance of one's identity is condition enough for us to achieve a moral judgement, an 'impartial' or 'universal' standpoint. As a result, when morally judging a social arrangement, we should judge it under the special circumstance of factual ignorance, which recommends (on behalf of our self-interest and for rationality's sake) that we adopt an assumption of **equal probability** of our occupying any particular position within that particular world.

> (...) an individual's preferences satisfy this requirement of impersonality if they indicate what social situation he would choose if he did not know what his personal position would be in the new situation chosen (and in any of its alternatives) but rather had an equal chance of obtaining any of the social positions existing in this situation, from the highest down to the lowest. (Harsanyi,1980:14)

This is Harsanyi's equiprobability postulate which he believes to be enough to guarantee the universality of our moral judgments. Harsanyi insists that this is human and humane morality. In terms of its consequences, this is tantamount to asking individuals to judge among different institutional arrangements according to their intrinsic merits. In that context, we expect them to recur to their moral preferences as defined, that is, preferences that do not depend on their particular roles inside the arrangements they are to judge, preferences formed from behind what John Rawls has called a 'veil of ignorance'.[6]

It now becomes clear which particular rational decision-rule is to apply under the circumstances of morality, which are circumstances of **uncertainty** (real or putative) concerning the individual's identity in the social arrangement.

> (...) an impersonal choice (preference) of this kind can in a technical sense be regarded as a choice between 'uncertain' prospects. (ibidem)

The individual performing the 'moral choice' will use as his decision-rule the principle of **expected** utility maximization, which is recommended as a principle of rational decision making under uncertainty, here (in the realm of ethics) as elsewhere (in the realm of decision theory). In particular, under the circumstances of morality which make appeal to the equiprobability model, as argued above, an individual **i** will be rational (in his moral judgments) whenever he chooses or judges among different social arrangements that which offers the greatest **average** utility level, W_i:

$$W_i = 1/n \sum_{j=1}^{n} U_j$$

where the U_j stands for all individual utility levels, and **n** is the number of the individuals. (The factual assumptions that interpersonal comparisons of utility are meaningful, and that the individual utilities are additive, are so far taken for granted).

---

[6] I should add the important difference that Rawls invites other conditions as also relevant to the original position apart from what is here called the 'self-interested' perspective. See chapter 6.

Thus, the utilitarian moral rule of the maximization of social utility is interpreted here as the maximization of **expected utility**, and, in particular along with the equiprobability model, as the maximization of **average utility**.

Harsanyi (1982) provides two kinds of justification for his utilitarian conclusion, one bearing on the 'veil of ignorance' assumption (the V-assumption, henceforth) and the other being essentially an axiomatic justification.

The first justification argues plainly enough that a moral judgment is a judgement from behind a veil of ignorance, for that is the natural deliverance from a personal or self-seeking bias that finds support in our common behavior. Without knowledge of our future position, though knowing some basic facts about social life and human psychology, it is natural that we place the same probability to the event of our occupying any particular position within a social arrangement when judging its relative acceptability. It is true that the full equiprobability assumption is only a fiction, but for an individual it can quite closely account for his serious effort to disregard his particular position and 'make his choice as if he thought he would have the same probability of taking the place of any particular individual in the society',[7] Harsanyi contends.

The idea of morality as a rational choice from behind a veil of ignorance has been attacked on many grounds. For one thing, different people might rationally have different attitudes to the same kind of life, Broome (1991) has argued. For another, Brian Barry has asked,[8] why bother to identify 'moral' judgments with those judgments made by someone trying to maximize **his own** prospects from behind a veil of ignorance? With yet another objection, Hammond (1991b) and Kelsey (1988) focus on the issue of attitudes to risk, the former arguing against the possibility that someone knows the different attitudes to risk that different people entertain, the latter pointing out the likely biases that we might expect to encounter in moral reasoning stemming from one's

---

[7] Cf. Harsanyi (1980), p.46.

[8] Cf. Broome, op. cit., quoting from B. Barry's Theories of Justice.

knowledge of one's own attitude to risk. In summary, there is much controversy surrounding the use of the V-assumption.[9]

Now, one argument in favor of the utilitarian principle seemingly different from the one relying upon the V-assumption alone is the so-called axiomatic argument. It recommends the utilitarian principle on account of rationality considerations alone in the context of 'moral choices', like the decisions concerning 'what the common interests of society are'.[10] What are the demands that consistency alone imposes on our moral judgements or preferences for them to be rational? According to Harsanyi these are to be found in the von Neumann-Morgenstern utility functions and Bayesian rationality.

What is peculiar about these functions that renders them so attractive as far as ethical considerations are concerned? Firstly, says Harsanyi, they are a representation of one's rational preferences when faced with risky choices; they are estimated in terms of one's behavior under

---

[9] We shall later concentrate on Rawls' criticism of Harsanyi's use of this postulate and his own alternative approach. It is worth mentioning here the Rawlsian defense of an alternative decision rule that would emerge from our rational choices from behind a veil of ignorance, as well as Harsanyi's reaction to that. This is the idea of maximin as the decision rule to be used by rational individuals in the original position, that is, that under uncertainty as to his own identity the individual should choose the action that offers the greatest prospect for the worst outcome. In contrast to Rawls, Harsanyi (1973,1980) regards the maximin principle as an irrational decision rule in the circumstances of the original position. His argument is quite straightforward: to suppose the maximin as the rational principle in the original position is equivalent to postulating probability one (or nearly so) to the person occupying the worst position in the social arrangement. But why should we? Harsanyi opts, instead, for the principle of expected utility maximization for he considers that the individuals are able to assign **probabilities** to the likelihood of their happening to fall under any particular place in the society, and also that, in the absence of information to the contrary, they shall assign **equal** probability to the occurrence of any such event. According to Rawls, however, it is not possible that individuals attribute objective probabilities to these events for they do not have any empirical evidence supporting these calculations. Rawls also argues against the Laplacian principle of insufficient reason that would recommend the equiprobability rule. Harsanyi replies that individuals assign **subjective and not objective** (empirical) or logical probabilities to the likelihood of these events. For him any rational decision maker, that is someone whose behavior is consistent with 'a few very compelling rationality postulates **cannot help** acting **as if** he used subjective probabilities. (More precisely, he cannot help acting **as if** he tried to maximize his expected utility, computed on the basis of some set of subjective probabilities.)' Cf. Harsanyi (1973, 1980), p.47, author's emphasis. In discussing the choice of a social arrangement in the original position, Harsanyi models it as a rational choice under uncertainty, as follows: "What decision rule would rational individuals use in the original position in deciding where a given set of institutions was or was not acceptable to them? In the terminology of modern decision theory, the initial position would be a situation of uncertainty because, by assumption, the participants would be uncertain about what their personal circumstances would be under any particular institutional framework to be agreed upon. There are two schools of thought about the decision rule to be used by a rational person under uncertainty. One proposes the **maximin principle** (...). The other - Bayesian - school of thought, which is now dominant, proposes **expected utility maximization** as decision rule under uncertainty. (Harsanyi,1980:38).

[10] Cf. Harsanyi (1982), p.48.

risk, and, therefore, they are expected to reflect people's attitudes to risk apart from the preferences themselves. Secondly, and more significantly, they also can be seen to represent, in terms of a cardinal utility measure, he claims, one's assignment of relative weights to the things one values in life, 'it express the subjective importance people attach to their various needs and interests'.[11] According to Harsanyi, this serves as a basis for the priority to be given to these claims in a conception of justice. In somewhat more technical language, the axiomatical justification of the utilitarian principle goes like this:[12]

> **Axiom 1: Individual rationality**. The personal preferences of all **n** individuals in society satisfy the Bayesian rationality postulates.
>
> **Axiom 2: Rationality of moral preferences**. The moral preferences of at least one individual, namely, individual **i**, satisfy the Bayesian rationality postulates.
>
> **Axiom 3: Pareto optimality**. Suppose that at least one individual **j** (j = 1,...,n) personally prefers alternative **A** to alternative **B**, and that no individual has an opposite personal preference. Then individual **i** will morally prefer alternative **A** over alternative **B**.

Commenting on the axioms, Harsanyi adds:

> Axiom 3 is a very weak and hardly objectionable moral postulate. Axiom 1 is a rather natural rationality requirement. Axiom 2 is an equally natural rationality requirement: in trying to decide what the common interests of society are, we should surely follow at least as high standards of rationality as we follow (Axiom 1) in looking after our own personal interests.

On the way to the utilitarian principle, the argument follows:

> Axiom 1 implies that the personal preferences of each individual **j** (j =1,...,n) can be represented by a von Neumann-Morgenstern (=vNM) utility function U_j. Axiom 2 implies that the moral preferences of individual **i** can be represented by a social welfare function W_i, which mathematically also has the nature of vNM utility function. Finally, the three axioms together imply the following theorem:

---

[11] Cf. Harsanyi (1980). In Harsanyi (1982) there is the following statement concerning the 'real' purpose of the vNM utility functions: "even though a person's vNM utility is always estimated in terms of his behavior under risk and uncertainty, the real purpose of this estimation is to obtain cardinal-utility measures for the relative personal importance he assigns to various economic (and noneconomic) alternatives."(p.53)

[12] I quote from Harsanyi (1982), p.48/49.

> **Theorem T**. The social welfare function W_i of individual i must be of the mathematical form:
>
> $W_i = a_j U_j$ with $a>0$ for $j=1,...,n$.
>
> This result can be strengthened by adding a fourth axiom:
>
> **Axiom 4: Symmetry**. The social-welfare function W_i is a symmetric function of all individual utilities. (That is, different individuals should be treated equally.)
>
> Using this axiom, we can conclude that
>
> $a_1=...=a_n>0$.

Harsanyi then concludes:

> the four axioms of the present section make only very weak philosophical assumptions. They should appeal to everybody who believes in Bayesian rationality, in Pareto optimality, and in equal treatment of all individuals. Yet these very weak axioms turn out to be sufficient to entail a utilitarian theory of morality. (Harsanyi, 1982:49)

As a result, 'the common interests of society' are solved in terms of a weighed sum of individuals' utilities, where the weights are equal. The practical consequence is that these interests conflate into those of the average individual, that is, with the average utility.

We now proceed to the discussion of some pieces of the axiomatic argument.[13] In the next two sections I turn to the so-called factual or knowledge assumptions on which the utilitarian principle thus restated is based, respectively, the knowledge of the individuals' utility functions in section 3, and the possibility of interpersonal utility comparisons in section 4. These are crucial knowledge assumptions as far as an **additive** social welfare function is concerned. In section 5, I discuss the symmetry axiom in relation to the equiprobability assumption.

Let us turn now to the issue of the knowledge of the individual utility function.

---

[13] For a thorough criticism of the use of the vNM utility functions to underscore the utilitarian morality, see Hampton (1994). As for the Pareto optimality as a 'natural' requirement, I recall Sen (1970) and Elster's (1982) criticisms, as well as Hardin's (1984). The former point out the frailty of a preference approach as a supporter of any view of autonomous choices; the latter points out the implicit and controversial distributive impact of the Pareto criterion.

3. The Knowledge Assumption of Individual Utilities: practical reason in terms of theoretical reason

To begin with, Harsanyi's utilitarianism stresses the role of 'preferences' instead of desires in the explanation of human rational behavior, as I have already remarked. In particular, he rejects the subjectivism implied in the hedonic version of classical utilitarianism and adopts a quite objective stance with respect to human motivations.

In a recent work, Harsanyi (1992) clarifies a little further his dissatisfaction with the classical view and also the need for expanding utility theory in order to accommodate ethical issues. As a result, he works out a preference-based theory of human motivation that includes a detailed normative argument for domain-restriction. The grounds for further restrictions on preferences, besides the usual consistency requirements, stem from the need to put them on a par with the idea of a 'social' (as opposed to an individual) dimension of utility. In the following we shall go through Harsanyi's initial steps towards the idea of social utility from his peculiar **rule**-utilitarian perspective.

The starting point is the assumption of 'preference autonomy', which states that what is good or bad for an individual depends only on his own desires and preferences **as these exist at a 'deep level'**.

In examining this assumption we should note that the rule-utilitarian standpoint demands that our actions be in line with a particular moral rule, applicable to the circumstances, and that this particular rule, in turn, be in line with the utilitarian moral rule of maximization of overall utility (or, for that matter, average utility). Therefore, in contradistinction to 'act-utilitarianism', the morality of our acts is not to be judged **directly** in terms of their effect on overall utility. In other words, the rule-utilitarian requires strict restrictions on individual preferences in the sense that actions are to be judged permissible according to whether they abide by a particular norm (or set of norms) applicable to the circumstances, and, through the particular norm, on account of its effects on overall utility.

So, our acts are to be judged against some standard of correctness that makes appeal, suggests Harsanyi, to a notion of what our 'real' or 'true' interests are as well as to what extent these

interests are genuine. For that matter, Harsanyi (1982) defends the exclusion of 'anti-social' preferences based on spurious motivations like sadism, envy, resentment and malice: these feelings of people 'cannot claim for a hearing when it comes to defining our concept of social utility'.[14]

As for the 'truth-content' of our interests, this will become clear as soon as we realize what is involved in the shift from desires to preferences in utility theory. Essentially, Harsanyi argues, the desires-approach fails to capture the fact that we are 'social creatures', which explains to an important extent the corollary fact that we have non-subjective motivations as well. In other words, we also have some objective goals, some outside accomplishments besides achieving desirable states of mind:

> By nature and by choice we are **social creatures** rather than solitary ones. Perhaps this is the deeper reason why we have genuine altruistic concern for other people; why we want to make worthwhile contributions to some objectives we share with many other people in intellectual, artistic, social and political life; and why our interests reach out far beyond our own inner experiences and very much extend to states of affairs in the outside world. (As I have already suggested. to pay close attention to the outside world is also clearly in our self-interest.) (Harsanyi, 1992:3/4)

This objectivity usually implicit in a broad preference-approach is strengthened by circumstances peculiar to matters of welfare and ethics, Harsanyi asserts. This means that in these fields we should like to distinguish the 'true' preferences and interests of a person from his ill-informed or badly grounded ones:

> in welfare economics and ethics we want to distinguish between those choices of a person that really express his true preferences and his true interests at a deeper level, from those choices of his that fail to do so because they are based on incorrect information or on ignorance or neglect of some important information. (ibidem:5/6)

So, whereas generally in economics all we need is some suitable consistency axioms that might render our preferences representable by utility functions, in welfare economics and ethics we want to constrain our preferences further, we would also like to know whether our preferences are **well-informed**. In addition, as we have already mentioned, Harsanyi asserts that we would like

---

[14] Cf. Harsanyi (1982), p.56.

to consider only those preferences that are **genuine**, and not spurious such as those springing from many forms of compulsive behavior and self-deception. Our welfare or ethical judgments should have as their raw material the personal preferences of individuals but only as far as these are informed-preferences (which might be in opposition to the individuals' actual, ill-informed preferences). Additionally, they should be genuine preferences. The relation between our informed and our actual preferences is as follows:

> A person's **informed** preferences and desires are defined as what his preferences and desires would be like under some hypothetical conditions. This of course means that they are **not** directly observable empirical variables as his **actual** preferences and desires are but rather are **theoretical constructs**. (ibidem:6)

So, in welfare economics and ethics, individual utility functions represent only the 'true' or informed-preferences in the above sense:

> It seems to me that, at least in welfare economics and ethics, a person's utility function should be defined in terms of his hypothetical **informed** preferences rather than in terms of his **actual** preferences because some of the latter may be badly mistaken. (ibidem:6/7)

How do we know the informed-preferences of a person? These can be identified as his actual preferences as freed from factual errors. All we can do, Harsanyi warns in a footnote, 'is to define i's informed-preferences as the preferences he would have if his beliefs about the relevant facts were correct **as judged by the best information I have** about these facts, which of course might contain some errors'

When discussing informed preferences with regard to the problem of drug-addiction, Harsanyi compares our informed preferences or true interests to our well-considered judgments on the matter, thus calling attention to the cognitive aspect of our motivations. Indeed, this issue is deepened in a brief commentary by Harsanyi on Hume's model of human motivation. There Harsanyi clarifies the role of reason (in the form of factual assumptions or well-grounded beliefs) in the determination of our motivation, in contrast to the Humean dictum that reason is but the servant of our passions:

> unlike Hume, I assume that our 'passions', i.e., our desires and preferences, themselves crucially depend on the factual assumptions suggested by our reason. It is not the 'activist' nature of our reason itself that establishes this relationship between our reason on the one hand and our desires and preferences on the other hand but it is rather the dependence of the latter two on such factual assumptions. (ibidem:17)
>
> (...)
>
> Moreover, whereas Hume wants to restrict our reason to suggesting **means** to our ends, themselves determined by our 'passions', my model assumes that our reason not only suggests **means** to our ends, on the basis of the **instrumentally** desirable attributes of various alternatives, but also suggests **ends** worth pursuing, on the basis of the **intrinsically** desirable attributes of these alternatives. (ibidem:17)

Therefore, in a sense, 'true' preferences are rational ones, and not only instrumentally, as a means to produce pleasure, but also intrinsically, as ends in themselves, for they spring from the suggestions (the factual assumptions) that reason provides. Here, the normative role of **theoretical** reason is suggested as a source of morality: whenever we disregard something which should be valuable to us, reason may indicate that thing to us. But, what should be valuable to us? Harsanyi, following Parfit, suggests a list of such things, that are our 'natural desires'.[15]

In addition, preferences express a view of our personal priorities when it comes to the satisfaction of these desires. So, the preferences-account may capture the fact that our preferences, actual and informed, vary a great deal among different people (although the basic desires are nearly the same):

> Their preferences may be very different between alternative ways of satisfying the **same** basic desires. For instance, they may prefer very different jobs. Moreover, they may have very different priorities in satisfying **different** basic desires. Thus, they may have different priorities in dividing their time between their job and their family. (ibidem:24)

---

[15] Cf. Harsanyi (1992), there follows a list of our basic desires ('even though with considerable interpersonal variation in intensity'): "desire for material comfort and for physical and economic security; for freedom to control their own lives; for having good health; for jobs suitable to their personal abilities and personal interests; for further developing their abilities; for deep personal relations in mutual love, in marriage and in true friendship; for having children and for being a good parent; for knowledge and for understanding the world and their own place in the world; for enjoyment of beauty in nature and in art; for having access to the ordinary pleasures of human life; for worthwhile accomplishments of some kind; and for making their own behavior consistent with their basic moral values." On reflection, are these basic desires not a set of **values** to some extent? Suppose they are not, then, the role of reason in indicating them to our 'will' is somehow impossible to fulfill.

That there are some such substantive goods intrinsically valuable to human beings is a consequence of the fact that we all share a **common human nature** and **common biological and psychological needs**. However, we also have distinct personal attitudes towards those goods. The preference approach can cope with both aspects. It gives room for a theoretical reason to suggest what our ends should be, in accordance with our ultimate natural ends. It is also sensitive to the different priorities reflecting different personal attitudes of people when trying to satisfy the same natural ends. Likely constraints embodied in this approach are, on the one hand, reason itself understood as a capacity of revealing the true connections between one's personal ends and one's ultimate (natural) ends, on the other hand, the consistency requirements for rational preferences as dictated by the canons of rational choice theory.

Harsanyi contends that in the end, we may evaluate a person's choices and preferences with the aid of the informed-preferences approach, as well as the norms of morality that specify other people's morally protected interests as these latter are dictated by rule-utilitarianism. These norms would eventually recommend an action on the ground of its impact over social utility.

To establish what one person's informed-preferences are we have to contrast his actual preferences with those of other knowledgeable people:

> suppose that most knowledgeable people assign a high utility to some benefit **A** because **they know from personal experience** or from what **they have learned about other people's experiences** that **A** tends to add a lot of extra satisfaction to one's life. Yet, a particular individual seems to have no interest at all in obtaining **A**. Then, it will be a reasonable assumption that if he were better informed then he would likewise assign a high utility to **A**, more or less within the same range as other people do - except if he has some special **disability** preventing him from taking full advantage of this benefit **A**. (For example, some people may have special psychological difficulties in making friends and in retaining them over any length of time. Other may be unable to enjoy some of the greatest works of world literature for lack of education, and so on.) (ibidem:30, my emphasis)

In this passage Harsanyi sets down some grounds for paternalism, as proper education and therapy are invoked as mechanisms for the correction of mistaken factual assumptions. We also have a hint of how these assumptions are inductively formed, i.e. from (some) people's direct or

indirect experience. However, what is missing is a view of how these inferences relate to the supposed objective or true belief.

In conclusion, a distinct view of preferences than the 'revealed preferences' approach is undertaken here. The 'revealed preferences' approach has vindicated the dismissal of the 'desires account' on the heuristic basis that a theory of what motivates an agent is not really necessary, it is indeed costly in 'heuristic currency', as far as the explanation of economic behavior is concerned. The Harsanyian informed-preferences approach builds, instead, upon the desires approach. It tries to say something about 'rational' desires as well as 'true' or 'rational' preferences; it ultimately accepts that there is some basic psychological state we want to foster and it only adds a 'social' dimension to our search for satisfaction. This 'social' dimension, however, is given as the assumption of a common nature and psychology which all human beings share, and which may be best known in the experience of some knowledgeable people. A view of how these experiences relate to our common nature and psychology and therefore how we come to recognize the truth-content of our preferences, although crucial, is however, absent.

4. The Knowledge Assumption of Interpersonal Comparisons

Apart from the possibility of knowing someone's informed-preferences, and thus granting an access to that person's utility function, a rule utilitarian would want to establish the possibility of making meaningful interpersonal comparisons of such utilities, for this is a crucial assumption as far as an additive social utility function and the utilitarian principle are concerned. That is, we can only add the utility functions of different people, therefore forming a social utility or social welfare function, if these functions are commensurable.

To begin with, according to Harsanyi, we cannot avoid making interpersonal comparisons 'if we want to make our moral decisions in a responsible manner'.[16] He goes on to state that making an interpersonal comparison 'between the utility levels of two individuals **i** and **j** amounts to asking

[16] Cf. Harsanyi (1992), p.31.

the question **how much satisfaction** each of them derives from his own objective position, given his own preferences and, more generally, given his own personal attitudes'.[17]

Indeed, the possibility of interpersonal comparisons of utility turns, first of all, on the ability to decompose the individuals' utility functions into two distinct sets, namely, the set of objective positions (the 'substantive goods', economic or otherwise) and the set of personal attitudes (biologically and biographically determined). It also depends on the strong assumption that these distinct personal attitudes may be incorporated by someone going through the appropriate thought experiment. And finally, it depends on there being an appropriate standpoint from which to compare the different levels of satisfaction in an adequate way.

Harsanyi (1992) suggests that there are two different perspectives that could be followed in making these comparisons. These are the **third-person perspective** and the **first-person perspective**.

According to the third-person perspective, we would want to infer the individuals' satisfaction levels from the **laws of human psychology**, and from a **knowledge** of each individual's **extended alternative**, that is, his pairs of objective positions and personal attitudes. Thus, in a 1973 article,[18] Harsanyi proposes the following justification for interpersonal utility comparisons:

> the ultimate logical basis for interpersonal comparisons (...) lies in the postulate that the **preferences and utility functions of all human individuals are governed by the same basic psychological laws**. My utility function may be very different from yours. But, since both of our utility functions are governed by the same basic psychological laws, if I had your personal characteristics - and, in particular, if I had your biological inheritance and had your life history behind me - then presumably I would now have a utility function exactly like yours. This means that any **inter**personal comparison I may try to make between your present utility level and my own, reduces to an **intra**-personal utility comparison between the utility level **I** myself **do** now enjoy, and the utility level **I** myself **would** enjoy under certain hypothetical conditions namely if I were placed in your physical, economic, and social position, and also had my own biological and biographical background replaced by yours.

---

[17] Cf. ibidem, p.31.

[18] Reprinted in Harsanyi (1980).

> This means that interpersonal utility comparisons have a completely specific theoretical meaning, in the sense that, 'under ideal conditions', i.e., if we had full knowledge of the psychological laws governing people's preferences and their utility functions, and also had sufficient information about other people's personal characteristics, then we could make perfectly error-free interpersonal utility comparisons. (Harsanyi,1973,1980:50/51)

Harsanyi argues that this possibility has its shortcomings. In particular, the fact that 'our understanding of these psychological laws is far from being sufficient for doing so'.[19]

He goes on to suggest the first-person perspective:

> we must use an alternative approach by taking a **first-person perspective** and by trying to achieve an **empathetic understanding** of what it may be like to be in either individual's objective position with the relevant individual's own personal attitudes. Moreover, in keeping with this first-person perspective, I have also suggested that each of us should ask himself whether **we ourselves would prefer** to be faced with the extended alternative (A_i,P_i) [for the objective positions and personal attitudes, respectively] or with the extended alternative (A_j,P_j). I have argued that in deciding which way our own preference would go between these two extended alternatives, we should concentrate on these **two alternatives themselves**, and should try to abstract from our own personal attitudes as much as we can. (ibidem:31/32)

As for the possibility of success in such undertaking, Harsanyi warns:

> Of course, I have realized that we can never get rid completely of our personal biases in making interpersonal comparisons of utility, but have been convinced that we can go a long way in doing so if we really try. (ibidem:32)

Still, some factual evidence seems to lend support to the possibility of meaningful interpersonal comparisons:

> When different people make interpersonal comparisons of utility between the same two individuals or between the same two social groups, they may perhaps arrive at somewhat different conclusions, but most of the time their conclusions will be

---

[19] Cf. Harsanyi (1992), p.31.

> close enough to show that making such comparisons is not an altogether hopeless undertaking. (ibidem:33)[20]

However, it may be contended that there is no clear-cut conceptual distinction between the first-person and the third-person perspectives, for the third person perspective actually invites us, first-persons, to take up the standpoint of an external ideal observer (Harsanyi's terms). That we cannot eventually fully succeed in achieving this standpoint is but an inevitable consequence of our actual limited knowledge, a state of affairs, by the way, that could be improved upon, he seems to believe. Besides, he asserts, there is plenty of evidence for agreement on the side of competent observers in such matters.

Binmore (1994), however, insists that we look at Harsanyi's contribution as twofold: on the one hand, Harsanyi lends support to the conception of an ideal observer whose interpersonal comparisons stem from a **sympathetic** understanding of everyone's positions; on the other hand, a later Harsanyi seems to stress a more inductive **empathetic** understanding, to which Binmore is frankly sympathetic.

The ideal-observer perspective is akin to what has been known as the 'Harsanyi doctrine', after Aumann's (1987) suggestion. This is the sympathetical ideal carried to perfection:

> if Peter had Paul's biological makeup, had Paul's life history behind him, and were currently subjected to Paul's environmental influences, then he would presumably have the **same** personal preferences as Paul. (cited in Binmore, 1994:61, quoted from Harsanyi)

Therefore, the ideal observer, being capable of performing such exchanges, appears as the aggregative procedure through which personal preferences are put together in a proper way in order to generate social preferences. Putting it somewhat differently, the social utility function is

---

[20] A possible way to achieving such comparisons is suggested with the use of Charles Taylor's theory of sympathetic understanding of the values of a foreign culture: "Taylor suggests that we must give a sympathetic hearing to these foreign values, retaining our own old values as much as we feel reasonable to do, perhaps with appropriate modifications, but being also willing to revise our own values in the light of these foreign values whenever this seems to be the proper course to take. By following this approach, in the end we may achieve a broader point of view that does justice both to our own values and to those of the foreign culture we are trying to understand. In trying to understand another person's attitudes and values, we must follow a similar approach." (ibidem:32/33)

the utility function of an ideal observer, whose preferences are an aggregate of the individuals' preferences in the society. He produces the scale on which everyone's utility functions are compared, for he knows and expresses the 'true' conversion ratios between the various individuals' utility scales. In this way, the interpersonal comparisons are solved as the intrapersonal comparisons of a rational individual after having performed the sympathetic experiment of putting himself in everyone else's shoes.

The ideal observer can then be any rational and well-informed individual:

> a rational individual will try to base his social welfare function on the 'true' conversion ratios between the various individuals' utility units (i.e., on the conversion ratios that would presumably be used by observers who had full information about these individuals' personal characteristics and about the general psychological laws governing human behavior). But if he does not have enough information to ascertain these 'true' conversion ratios, then he will use his best **estimates** of the latter. (Binmore,1994:62)

That a rational individual will converge when assessing his situation to the individuals actually happening to fall under the various positions in the social arrangement will only happen if we adhere to an assumption of convergence of thought processes, Binmore warns. But why should we?

> If a person's **state of mind** is understood to include each nuance of attitude and shade of opinion, then it is surely impossible for one person ever to be certain of the state of mind with which another person approaches a problem. (Binmore,1994:213)

Additionally, we cannot be certain whether a particular small difference is irrelevant or instead may render someone's thought process quite unpredictable.

A more convincing 'non-teleological' view of interpersonal comparisons, according to Binmore, would be based on the possibility of **empathetic** (the first-person perspective) rather than sympathetic preferences. This is also justified by the work of Harsanyi, he points out. It is the view that

> to make sense of comparisons of utility across individuals, it is necessary to go beyond the **personal** preferences (...) One also needs to take account of the **empathetic** preferences (...), the 'extended sympathy preferences'. (ibidem:283)

It will be helpful to distinguish between sympathy and empathy. In the sympathy approach, Binmore argues, a person happens to identify so strongly with another that he is unable to separate his interests from that person's. In the more appealing empathetic case, he contends, a person may still identify with another without losing sight of his own preferences, that is, he does not cease to separate his own preferences from the others'. This latter kind of identification is more cogent in Binmore's view.

Besides, when making fairness judgments, he adds, we need more than empathetic identification; we also need to develop empathetic **preferences**. 'Eve must be able to say **how much** better or worse she feels when identifying with Adam than when identifying with herself'. In order for Eve to do so she must be equipped with empathetic preferences: 'I am expressing an empathetic preference', says Binmore, 'when I say that I would rather be Eve eating an apple than Adam wearing a fig leave'.[21]

Suppose, as Binmore suggests, that some individual **i** is morally evaluating a social arrangement **C**. Since he does not know whether he will be Eve or Adam within this arrangement, according to Harsanyi he should follow the equiprobability model and assign equal probabilities to these two events. He has then two possible outcomes consisting of the outcomes associated with Eve's and Adam's positions in such a social arrangement, each available with probability 1/2. His situation can be represented as an expected von Neumann and Morgenstern utility function:

$$W_i(C)=1/2\ \{v_i(C,A) + v_i(C,E)\},$$

where $v_i(C,A)$ stands for the utility individual **i** associates to his being Adam in the social contract **C**, and $v_i(C,E)$ stands for the utility he assigns to his being Eve in **C**.

---

[21] Cf. Binmore (1994), p.290.

There is nothing to guarantee that people will arrive at the same empathetic preferences, which is a condition we would like to see met. So, Binmore asks, under what conditions are these value judgments to be the same for everybody in the society? He himself sides with a view that social evolution will tend to produce an agreement in these judgments.[22]

Harsanyi, however, states that a rough agreement is already in existence. However, if he insists on this line, he will miss the distance between this fact and the norm of morality he has been cautiously trying to construct: **ought** what exists to exist? In other words, **should** the agreement that exists exist, and if so, then why? On closer inspection, Harsanyi lets these empirical judgements be made by competent observers and he believes that they would agree on a scale to be used for comparisons. But, again, why take any such agreement as evidence of 'truth' (or convergence to 'truth') about interpersonal comparisons?

Summing up what we have arrived at so far, a 'sympathetic' approach sticks to a view that interpersonal comparisons are possible thanks to the so-called Harsanyi doctrine or, in his own words, the 'similarity postulate'.[23] The meaningful IC are those arrived at through the standpoint of an **ideal** observer. The similarity postulate can be restated as a view that thought processes are convergent, that is, placed in the same situation people will reason in the same way therefore leading to the same results. In addition, in morally assessing a social arrangement, people are taken to form their sympathetic preferences from behind a veil of ignorance. So, apart from being thought to come up with the same interpersonal comparisons, people are supposed to abide by the equiprobability model which asserts that under uncertainty (or ignorance) of their actual identities people assign the same probability of their happening to occupy any of the possibly

---

[22] On this interpretation, it seems to me, we have a mixing between the utilitarian and the evolutionary approaches where the overall happiness emerges as an *unpredicted* consequence of an individual rational choice under uncertainty. We may, however, ask: how can anyone know that in advance?

[23] Cf. Harsanyi (1982); he explains his 'similarity postulate' in terms of imaginative sympathy: 'We imagine ourselves to be in the shoes of another person, and ask ourselves the question, If I were now really in **his** position, and had **his** taste, **his** education, **his** social background, **his** cultural values, and **his** psychological make-up, then what would now be **my** preferences between various alternatives, and how much satisfaction or dissatisfaction would I derive from any given alternative? (An 'alternative' here stands for a given bundle of economic commodities plus a given position with respect to various noneconomic variables, such as health, social status, job situation, family situation, etc.). In other words, any interpersonal utility comparison is based on what I will call the **similarity postulate**, to be defined as the assumption that, once proper allowances have been made for the empirically given differences in taste, education, etc., between me and another person, then it is reasonable for me to assume that our basic psychological reactions to any given alternative will be otherwise much the same.'(p.50)

existing positions within a social arrangement. The 'sympathy' view bears on the postulates of Harsanyi doctrine and of equiprobability, the former providing grounds for interpersonal comparisons, and the latter translating the utilitarian moral principle of 'maximization of social utility', into the 'maximization of mean utility'. Two over strong assumptions of the argument deserve mention: that thought processes are convergent, and that under uncertainty people tend to assign equal probability to the various events.

An 'empathetic' perspective, on the other hand, starts from an assumption that people are capable of performing empathetic identification and forming empathetic preferences, which though not perfectly identical, tend in practice to be close enough to provide a justification for the utilitarian principle. It does not turn, at least directly, on the Harsanyi doctrine, but still keeps the equiprobability postulate. Still, in this case, how can we know whether the actual intersubjective agreement is correct? We lack an objective standard, which seems to be necessary if we want to utter 'ought-statements', not derived from 'is-statements', alone. Actually, this objective standpoint is sometimes identified with that of a **competent observer,** where the meaningful IC are those arrived at by some 'more knowledgeable people'. Again, a view of how the knowledge held by those people relates to the truth-content of IC is lacking.

In view of the preceding, the two approaches do not appear that different, revealing instead a common way of dealing with 'mistaken factual assumptions.' The empathetic approach may be seen as a special case of the more general rule (worked out by supporters of the Harsanyi doctrine, i.e., Aumann (1976)[24]) that with different information people end up with different beliefs; these beliefs would be the same were people to be fed the same information. They would then think in precisely the same way. With this in mind, empathy seems to be the best proxy for sympathy.

[24] Refer to chapter 3, on game theory, where I discuss the Harsanyi doctrine.

5. The equiprobability model, the principle of insufficient reason and the symmetry among people

We now turn to Harsanyi's proposed relation between his equiprobability postulate, the principle of insufficient reason, and the symmetry axiom. It will emerge that the utilitarian morality is not so much firmly based upon an axiomatic for rational choices as upon a rather badly argued moral assumption.

The idea that 'equal probabilities should be assigned to different events unless some reason can be found for distinguishing between them'[25]is usually credited to Pierre Laplace's 1713 work. However, it actually goes back to a work by Jacob Bernoulli (1654-1705) the first formulation of the principle of insufficient reason (IR-principle), as Luce and Raiffa (L&R) report. It runs like this:

> if there is no evidence leading one to believe that one event from an exhaustive set of mutually exclusive events is more likely to occur than another, then the events should be judged equally probable. (Luce and Raiffa,1957,1964:284)

The vagueness of this definition has given rise to a number of qualifications, but, L&R argue, these have not settled three important difficulties: one concerning the **listing** of the states, for there will be in numerous cases the possibility of different descriptions of the states of nature; another, which is connected to the former, referring to the meaning of **equally likely**, for it is not independent of the enumeration of the states itself, therefore implying a circularity in the criterion; a third relating to the difficulty of enumerating the states when 'there are an infinite set of pertinent states of nature'.[26]

In spite of these analytical difficulties, the IR-principle appears to be the reasoning supporting the Harsanyian equiprobability postulate. And this, in turn, is to be given an ethical connotation. To begin with, how can this postulate, whose logical roots in the IR-principle are quite clear, be given an **ethical** meaning? Let us quote Harsanyi in search for clarification:

---

[25] Cf. Binmore (1994), p.305.

[26] Cf. Luce and Raiffa (1957,1964), p.285. Of course, as Alan Kirman pointed out to me, equal probabilities may be meaningful over an infinite non-countable set, as for example the unit interval. However, the objection raised by L&R seems to be sound when it comes to an indefinite set.

> My equiprobability assumption obviously can be regarded as an application of the principle of indifference [the Laplacian principle]. But it also has another possible interpretation. It may be regarded as an expression of the purely **moral** principle that, in making basic moral value judgments, we must give the same **a priori** weight to the interests of all members of the society. (Harsanyi, 1980:63,fn10)

Being proposed in this way as a principle that by attributing equal probabilities (the uniform probability argument) to the events called 'personal' or 'social positions', thereby showing equal respect for persons and positions (the symmetry argument), the utilitarian principle emerges as the humanistic moral code par excellence:

> Moral philosophy can point out the fact of fundamental importance that in ultimate analysis all non-humanistic codes of behavior are merely expressions of contingent personal preferences - though possibly of very disinterested preferences - on the part of the people adopting these codes: whereas the code of impartially sympathetic humanism is the only one which by definition gives the same **equal weight** to the preferences of any other person as well. (This last statement, of course, presupposes the possibility of making operationally meaningful interpersonal utility comparisons.) (Harsanyi,1958,1980:35)

The defense of the equiprobability model, then, seems to rely on the warrant this model yields as to the non-personal character of one's moral judgments; to assume equal probabilities is to provide a mechanism whereby personal considerations would play no relevant role in a person's assessment of a social arrangement. Moreover, this is supposed to be tantamount to securing different people and positions an equal treatment. The argument is twofold: on the one hand, we want an impartial judgment about the relative merits of different social arrangements, one that secures equal treatment to everyone's preferences; on the other hand, we recognize that this achievement is only successful in the event that we neutralize the impact of 'contingent personal preferences'. Is this enough on the way to 'impartiality'?

There are, one might object, other relevant considerations on the way to impartiality, beside the neutralization of personal preferences. To begin with, we might similarly want to constrain 'altruism'.[27] We may want to limit the values of people not just their interests, that is to say, the

---

[27] See chapter 1, for some problematic issues related to altruism.

different personal views of the common interests and good of society, to some standard of universality (as in Rawls). Certainly, Harsanyi makes room for genuine altruism when he includes objective outside accomplishments and values as possible objects of people's 'personal' (as opposed to 'moral') preferences, so that what a person wants may be not only selfish pleasure but also the pleasure or satisfaction that comes to other people. In this sense, the neutralization of personal preferences would also encompass altruism.

Still, one may also want to accommodate a concern about 'equality' into one's moral reasoning. For that matter, the IR-principle is claimed to make an appeal to one's concern about equality, for it recommends that equal weights be assigned to the preferences of everyone in the Harsanyian interpretation. In this way, symmetry would stand for equality. On reflection, however, equal probabilities are not applied to different people's lives **as they themselves see these lives**; these were already calibrated by a view of what their well-informed and genuine preferences should be before probabilities are assigned to them. Of course, we might conceive of people's agreeing with the calibration once sound objective arguments for it are offered; on this matter, however, there remains the pending problem of a unsatisfactorily grounded intersubjective agreement as discussed in section 3. Moreover, the additive form of the utilitarian principle presupposes that interpersonal comparisons have already been made so that the individuals' utility functions were already given different weights from the ideal observer's scale, a point we have pursued in section 4. Therefore, the assignment of equal weights, in the end, is equivalent to stating that social utility is a mere sum of people's utility functions as corrected by appropriate factual assumptions and cleaned of spurious motivations as well as put on a par by the ideal observer's IC. Turning to our previous remarks on the difficulties in handling the principle of insufficient reason, the problems of 'listing' and 'giving meaning' to the 'equally likely' condition stand out strikingly here. The equiprobability model cannot be claimed to express our genuine concern for equality: the alternatives themselves (the individual utility functions and their weights) were 'listed' so as to match one's prior conception of what should be 'equally likely'.

As a final remark, it is worth noticing that the IR-principle gives 'Harsanyi rule-utilitarianism' a different flavor than the pure subjectivist probability view he claims to adopt. To be sure, Harsanyi

seems to see the logical probabilities as a special case of the broader subjective probabilities framework: logical probabilities are subjective probabilities governed by a principle of symmetry:

> (...) by logical probabilities I mean subjective probabilities completely determined by symmetry considerations (if appropriate symmetry postulates are added to the standard postulates of Bayesian theory)." (ibidem:63,fn11)

In this view, the twin approaches to the utilitarian principle, the V-assumption and the axiomatic, are not independent from each other.

6. Conclusion

The utilitarian theorem of maximization of social utility has been restated by Harsanyi in terms of expected (truly, mean) utility maximization. On different occasions Harsanyi justifies this principle in terms of the moral preferences of a rational individual or his social welfare function, where, that is, he takes a thoroughly impersonal or impartial stand. The aggregative procedure proposed, which seemingly skips Arrow's stylized impossibility, is the ideal observer who is able to know the right IC of preferences.

The moral preferences expressed in the individual's social welfare function can be known by modelling his choice as a rational choice under uncertainty where this uncertainty refers to his ignorance (real or putative) regarding the basic qualities and resources that he happens to possess. What kinds of arrangement would the individual favor as a result of this exercise? He would choose the one that offered the greatest average utility, Harsanyi argues. In reality, two arguments are given in favor of the utilitarian principle, the V-assumption argument and the axiomatic one.

The argument stemming from the V-assumption runs like this: since the individual does not know his position in each possible social state he will be inclined to attribute equal probabilities to his falling on each possible position and then choose that social arrangement which offers the prospect of greater mean utility. The axiomatic argument deriving from the requirements of rational 'moral' choices establishes the utilitarian principle via Bayesian rationality, Pareto optimality, plus a symmetry axiom. But, then we learn that the equiprobability argument (the V-

assumption) is a possible extension of the axiomatic argument, where logical (equal) probabilities are only subjective probabilities complemented with symmetry considerations.

Symmetry, then, is at the core of the argument. It appeals to our equality concerns: people's preferences should be treated equally, Harsanyi declares. Closer inspection however reveals that different personal or group preferences had already passed through the riddle of 'preference autonomy' and been given proper weights before they were considered equally. So, the core of the argument shifts to the knowledge assumptions that assume that there are true preferences (non-spurious and well-informed) as well as true IC. The argument now becomes either inductive or authoritative. Since there is no guarantee that people's ordinary experiences will necessarily converge to true propositions as the inductive version wants us to believe, the authoritative branch appears more promising. Yet, if we are to let some more knowledgeable people find out what is better for us, why bother to have our agreement thus far?

Chapter 6

ORDER AS JUSTICE (II): JOHN RAWLS' CONTRACTARIANISM

1. Introduction

Our concern in this chapter shall be exclusively with Rawls' view of social order and the way he relates it to a conception of justice. Rawls shall constitute our third example of models of the kind suggested in cell II of the matrix shown in chapter 1, where, that is, parametric-rational individuals are to choose the social world most to their liking. This, in (early) Rawlsian terms, is translated into a constrained individual rational choice of public principles of justice.

In fact, Rawls portrays the rationality of individuals choosing principles of justice in the following way:

> a rational person is thought to have a coherent set of preferences between the options open to him. He ranks these options according to how well they further his purposes; he follows the plan which will satisfy more of his desires than less, and which has the greater chance of being successfully executed. (A Theory of Justice: 143)

Moreover, these individuals are supposed not to have an interest in others' interests. However, they are not necessarily narrow egoists, but are endowed with an 'interest **of** a self'. Furthermore, individuals are not motivated by strategic considerations, they are not 'concerned to win but to get as many points as possible judged by their own system of ends'.[1] For our purposes in this Part, this suffices to characterize their rationality as parametric.

In addition, it is characteristic of this approach that the individuals are 'theoretically defined entities'[2] in accordance with a peculiar interpretation of the contractarian tradition, thus relying little or not at all on any view of human nature. Individuals are seen as **constructed viewpoints**

---

[1] Cf. Rawls, 1971, p.145.

[2] Cf. Rawls, 1971.

that owe some consideration to social life, on both instrumental and intrinsic accounts. In relation to this, and in contrast to rule-utilitarianism, the role of general facts about human behavior is circumvented by certain **moral conditions** one should impose on human choices, in particular choice of principles of justice, to render them justifiable in the eyes of everyone at any time.

The conception of justice that emerges from this morally constrained rational choice should then specify not so much the configuration of a particular social state, but how the society, in the form of its major institutions is to assign basic rights and duties, as well as distribute the advantages of social cooperation among its members. It should therefore indicate what social arrangements might be considered acceptable.

Rawls attempts to show that proper reasoning (some suitable intrapersonal deliberation) may lead to the rejection of the major contemporary alternative conceptions based on utilitarianism and intuitionism, as well as perfectionism, and hopefully reach his conception of 'justice as fairness'. 'Justice as fairness' is simply the view that justice is a matter of proper justification of the choice of principles (which express a certain criterion for interpersonal comparisons) to the widest possible audience; it should want to articulate the fair situation for the principles to be chosen.

The Rawlsian principles of justice are defended with appeal to our considered judgments 'duly pruned and adjusted' as well as with the assistance of the canons of rational choice, for the choice of principles of justice may also be understood as constrained decision making under uncertainty. In addition, a certain 'vision' is also advanced by Rawls in support of the principles recommended by 'justice as fairness'. Some of these perspectives will be pursued in what follows.

Our purpose here is to examine the way the proposed relation between social unity, justice and interpersonal comparisons is worked out throughout significant pieces of Rawls' work.[3] To this

---

[3] Of course, the implicit assumption here is that Rawls' work has undergone substantial change in recent years. Ever since A Theory of Justice (TJ, henceforth), his major 1971 work, Rawls has not considered his arguments for 'justice as fairness' terminated. In the 70s and the 80s he wrote a number of articles and lectures where he qualified in many respects his earlier undertaking. In 1993 he eventually published a second *tour de force*, Political Liberalism (PL, henceforth), where the new developments of his thinking are put together and an effort to give them unity is undertaken. One of the most impressive shifts in his thought concerns the central role that the question of pluralism or 'reasonable pluralism' is to play, to the detriment of the emphasis that distributive issues had acquired in the context of the TJ. The

(continued...)

end, the views expressed by Rawls in a number of articles and, specially, in the TJ are pursued as they bring to light assorted arguments for that relation.

Provisionally, Rawls' view can be summarized thus: a society is well-ordered when the balance of 'reasons' among the individuals that pertain to it have a prior solution in the balance of 'reasons' within a properly defined individual, as far as a rule of adjudication of competing demands among citizens in the society is concerned. 'Social order' thus envisioned is a construction that springs from another construction which is the individual as a certain 'person', bearing on a certain hierarchy of reasons. An intermediary conception is designed to make explicit the constraints that this person is to be subjected to when deliberating. We say, then, that a criterion for interpersonal comparisons is made dependent upon an intrapersonal deliberation under certain constraints; the precedence of a 'moral' person over a 'bare self'[4] accounts for the possibility of the intrapersonal hierarchy of reasons.

From the preceding chapters, it appears that the straightforward understanding of social order in terms of rational individuals' motivations, for the normative purpose of designing a criterion for social choices, faces a number of difficulties, and has done ever since the Arrow result came to light. On the way to a solution, the IC, it has appeared, could not possibly be introduced into the analysis without further reflection, in particular because of the special incidence of value

---

[3](...continued)
radical message of the TJ lost a lot as a consequence of Rawls' looking for a wider basis of consensus to his theoretical construction that could encompass the most distant criticisms. The more intellectually tolerant Rawls was then accused of relativism, of proposing an unacceptable 'modus vivendi' when he was attempting to articulate the common ground of deeply divided (although reasonable) political, religious and philosophical doctrines that compound the mosaic of western democratic public culture, whose values at a deeper level his 'justice as fairness' wanted to express. On this, see Baynes, 1992. What is more, he was also charged with failing in that enterprise for he was not able to forsake a partisan attitude towards that very division (see Kukathas, 1990). As for his distributive message, critics complain that the demands of an overlapping consensus admittedly could be more easily met by the lesser and rather trivial (at least as Rawls himself considers, in the context of the TJ) social-justice target of equal basic rights and a social minimum rather than the more stringent difference principle. Cf. Rawls (1993a): "Whether the constitutional essentials covering the basic freedoms are satisfied is more or less visible on the face of constitutional arrangements and how these can be seen to work in practice. But whether the aims of the principles covering social and economic inequalities are realized is far more difficult to ascertain. These matters are nearly always open to wide differences of reasonable opinion..."(229) And: "(...)freedom of movement and free choice of occupation and a social minimum covering citizens' basic needs count as constitutional essentials while the principle of fair opportunity and the difference principle do not."(230). However, those who had rejoiced with the way Rawls' argument for the difference principle seemed to have properly addressed the question of the impact of arbitrary contingencies (social or whatever) in the distribution of resources in society felt justifiably frustrated with his withdrawal. See Baynes (1992).

[4] Cf. Rawls (1982), "Social Unity and Primary Goods".

judgments that seem to be attached to their incorporation. Harsanyi has provided a justification and a basis for interpersonal comparisons whose major liability lies in the cognitive assumptions adopted. Rawls wants to ground interpersonal comparisons in objective values.[5] To this end, he deepens the connection between morality and uncertainty in his account of the 'initial situation' the individuals face when choosing principles of justice, as if somehow trying to evoke deeper layers that ground their identities as rational individuals - those that should count when it comes to selecting principles of justice for regulating the basic institutions that are to preside their cooperation. The implication seems to be that the less we know materially about ourselves, the greater the prospect that we find some common grounds with others whom we are compelled by circumstances to interact with. Or, putting it differently, the lesser the contingent **knowledge** we have about ourselves, the more we have to rely upon a form of understanding, of **self-understanding** as some objective 'entity', to find our way.

I will concentrate on two distinct approaches Rawls has taken to accomplish this latter task, an earlier being deductive reasoning properly constrained, and a later one, the articulation of principles that seem to be latent in a public culture. In the course of our investigation, it has emerged, however, that the former argument falls down in the face of the veil of ignorance restriction, while the latter task falls down in the face of the 'publicity condition', or the extent to which we may speak interestingly of a 'public' reason.

The organization of the subsequent sections is as follows. In the next section, an overview of 'justice as fairness' is offered and special emphasis is given to the connection between social unity and justice. 'Justice as fairness' is presented as proposing a criterion for legitimate interpersonal comparisons in terms of an index of primary goods, which, it is claimed, renders social unity quite reliable.

---

[5] I quote here Rawls' statement on the IC issue: "Simply because we do in fact make what we call interpersonal comparisons of well-being does not mean that we understand the basis of these comparisons or that we should accept them as sound. To settle these matters we need to give an account of these judgments, to set out the criteria that underlie them. For questions of justice we should try to find some objective grounds for these comparisons, ones that men can recognize and agree to. At the present time, there appears to be no satisfactory answer to these difficulties from a utilitarian point of view. Therefore it seems that, for the time being at least, the principle of utility makes such heavy demands on our ability to estimate the balance of advantages that it defines at best an ambiguous court of appeal for questions of justice."(TJ:90/91)

Sections 3 and 4 then undertake to explore the connection between interpersonal comparisons and intrapersonal deliberation as the Rawlsian arguments suggest, trying to address the question why (and which) IC are legitimate. In section 3, the deductive argument of A Theory of Justice for this connection is examined, with special attention to the role of ignorance in the devising of the principles, where these latter are the outcome of a constrained rational choice under uncertainty undertaken by the parties to the agreement. In section 4, a different argument is set out where stress is put on a conception of the person, starting from an idea that citizens in a well-ordered society view themselves as free and equal moral persons. The ideas of freedom and equality that are embedded in our public culture are expressed in the principles, it is then claimed

I conclude, in section 5, with an outline of some of the difficulties with the arguments offered in sections 3 and 4, with a view to emphasizing a number of aspects (in particular, the unexpected importance of a theoretical reason in the agreement to the principles) that will prove relevant when it comes to comparing Part II with the remaining Parts of the thesis.

2. Order as Justice: an overview

According to Rawls, the best conception of justice should be judged in terms of its broader consequences to social unity, that is, if it fosters the many qualities one wants to see in one's community, namely, coordination of everyone's expectations, efficient allocation of community's resources, and stability of the social arrangement. We quote from the first paragraph of the Theory:

> In the absence of a certain measure of agreement on what is just and unjust, it is clearly more difficult for individuals to coordinate their plans efficiently in order to insure that mutually beneficial arrangements are maintained. Distrust and resentment corrode the ties of civility, and suspicion and hostility tempt men to act in ways they would otherwise avoid. So while the distinctive role of conceptions of justice is to specify basic rights and duties and to determine the appropriate distributive shares, the way in which a conception does this is bound to affect the problems of efficiency, coordination and stability.(...) one conception of justice is preferable to another when its broader consequences are more desirable. (TJ 6)

So, overall consequences matter greatly when it comes to the choice of a conception of justice. However, this does not make Rawls a consequentialist, for though it is obvious that consequences

do matter, he says,[6] one is not forced, simply by agreeing with this proposition into accepting a view that what is right is that which maximizes the consequences or 'the good', a view typical of consequentialist reasoning. It is of great importance, he argues, to know and assess the ways in which a conception of justice brings out desirable consequences.

The assumption here seems to be that a conception of justice stands out among the necessary conditions for the social unity as the condition of possibility for the other requirements (viz. efficiency, coordination and stability) to be fully effective, because it has a certain connection with people's motivation to trustfully adhere to a social arrangement in the first place. We want a conception of justice to attend to our concerns regarding the way a desired consequence, namely, the social arrangement, is brought about.

This latter idea becomes even clearer when one figures out who is the subject of a conception of justice, namely, the basic structure of society or its major political, economic and social institutions. A conception of justice specifies the way the public system of rules embodied in or identified with these institutions determine the assignment of rights and duties as well as the distribution of the benefits from social cooperation. The way a conception of justice does this is of fundamental importance for the effectiveness of social unity, claims Rawls. The existence of such a system, he notes, may be responsible for the desirable unintended consequences of the interaction of many individuals (a view, he argues, Adam Smith suggested in the Wealth of Nations by resorting to the image of an 'invisible hand'), as well as the possibility of strategic interaction to arise among people in society. In this regard, Rawls' view of social interaction is not at odds with a variety of social theoretical images of that interaction, only he devotes himself to

---

[6] In this regard, he notes that all "ethical doctrines worth our attention take consequences into account in judging rightness. One which did not would simply be irrational, crazy." (TJ:30)

articulate a fundamental premise of these images.[7] Without a well-grounded conception of justice, the desirable consequences of various interactions among individuals would be far less certain.

In this connection, Rawls remarks that 'although a society is a cooperative venture for mutual advantage',

> it is typically marked by a conflict as well as by an identity of interests. ... A set of principles is required for choosing among the various social arrangements which determine this division of advantages and for underwriting an agreement on the proper distributive shares. These principles are the principles of justice: they provide a way of assigning rights and duties in the basic institutions of society and they define the appropriate distribution of the benefits and burdens of social cooperation. (TJ:4)

We therefore would like our society to be well-ordered:

> a society is well-ordered when it is not only designed to advance the good of its members but when it is also **effectively regulated by a public conception of justice**. (TJ:4/5, my emphasis)

Which means that

> (1) everyone accepts and knows that the other accept the same principles of justice, and (2) the basic social institutions generally satisfy and are generally known to satisfy these principles. (TJ:5)

That the acceptance of the principles by individuals is common knowledge, as well as the fact that they effectively regulate the basic structure of society (and are known to do so), is made a crucial condition for **trust**. This, in turn, is deemed to be the cement of the social arrangement.

---

[7] Only Hayek would deny, on epistemological grounds, that this system of rules could be fully articulated. See Part III for a detailed exposition of Hayek's general argument. We should, however, also note here a remarkable similarity between Rawls and Hayek in terms of the way the former describes the system of rules as a complex system. He says: "We may also distinguish between a single rule (or group of rules), an institution (or a major part thereof), and the basic structure of the social system as a whole. The reason for doing this is that one or several rules of an arrangement may be unjust although the social system as a whole is not. There is the possibility not only that single rules and institutions are not by themselves sufficiently important but that within the structure of an institution or social system one apparent injustice compensates for another. The whole is less unjust than it would be if it contained but one of the unjust parts. Further, it is conceivable that a social system may be unjust even though none of its institutions are unjust taken separately: the injustice is a consequence of how they are combined together into a single system. One institution may encourage and appear to justify expectations which are denied or ignored by another."(TJ:57)

We need, then, to find grounds for agreement, the basis for a consensus on the conception of justice that could be reliably accepted by everyone in the society, and that could be effective as far as its major institutions are concerned.

> In this case while man may put forth excessive demands on one another, they nevertheless acknowledge a common point of view from which their claims may be adjudicated. (TJ:5)

The nature of the agreement is by now clear enough: it refers to the settling down of principles of justice for the basic structure of society once and for all, principles everyone could understand, accept and abide by. The procedure, then, follows the contractarian tradition. In fact, Rawls proposes the theory of justice as a particular chapter of the theory of rational choice, following the steps of earlier contractarian thinking. 'Justice as Fairness' is then proposed as a particular view of justice,

> as principles that free and rational persons concerned to further their own interests would accept in an initial position of equality as defining the fundamental terms of their association. These principles are to regulate all further agreements; they specify the kinds of social cooperation that can be entered into and the forms of government that can be established. (TJ:11)

The argument is proposed in terms of a procedure of 'construction'. It amounts to deriving a theoretically constructed well-ordered society from constructed individuals specially circumstanced, as the constraints represented in an initial situation are there to suggest.[8] The features of society, as well as the individuals and their constrained choice, are intuitively selected, there is no way to avoid intuition, Rawls declares. However he expects his argument intellectually to persuade and also he expects the principles of justice that are to emerge from the rational choice of such individuals to match our considered convictions in reflective equilibrium (they can be supported by a comparison between our considered judgments and the practical impact of the principles in terms of the specific policies they would favor).

Bringing the matter closer to home, in his 1982 article, "Social Unity and Primary Goods" (SU, henceforth), Rawls explores in a quite explicit manner the connection between justice, social unity

---

[8] See "Kantian Constructivism in Moral Theory", 1980 (KC, henceforth), for a further development of the 'constructive' project.

and interpersonal comparisons and he notes that justice enables social unity in that it establishes a criterion for comparisons between people. It establishes principles for evaluating distributive arrangements in terms of how well they cater to the individuals **proper** needs. The nature of the principles are further clarified: we need to discover principles that everyone can accept and that reflect a common understanding of what are the **legitimate claims** as well as the **legitimate ways** of advancing them in society. In contrast to utilitarianism, Rawls claims, an agreement on these principles needs only reflect a partial identification between people, enough to support the social unity, and not a complete identification as suggested by the idea of one 'rational good' to be pursued, such as happiness.

So, in contractarian terms, we need to set up an enabling initial situation, that is, a state of affairs conducive to the terms of a fair agreement. The traditional conception of the state of nature of contractarian thinking is, according to Rawls, of no direct use here, for it is rather a no-agreement point, of generalized egoism and, in this sense providing no ways out, if any, to the problem of reducing the likely conflicts that one should expect to emerge in such a form of interaction. On the other hand, Rawls points out, one should expect as a result of one's reasoning to a conception of justice to consider that this conception be general in form, as well as universal in application, that it be a **public** conception and that it be a 'final court of appeal in practical reasoning'.[9] What is then the 'appropriate status quo' for these principles to be reached, given these formal constraints and the rejection of the early-contractarian 'state of nature'?

First of all, one needs to distinguish the so-called 'circumstances of justice', for much of the answer turns on the way they are tailored. These are the circumstances where principles of justice apply and, therefore, characterize the choice problem that individuals would be facing:

> the normal conditions under which human cooperation is both possible and necessary (...) although a society is a cooperative venture for mutual advantage, it is typically marked by a conflict as well as an identity of interests. (TJ:126)

[9] Cf. TJ, p.135.

Building upon Hume's definition, the Theory selects 'moderate scarcity' as an objective circumstance of justice, as well as the subjective circumstance that the parties are 'mutually disinterested', in order to characterize the justice-wanting set.

Moderate scarcity means that

> While mutually advantageous arrangements are feasible, the benefits they yield fall short of the demands men put forward. (TJ:127)

Stated in this way, the objective circumstance stresses the fact that **any** advantageous arrangement likely to spring from an identity of interests would entail a distributive conflict.

That the parties display mutual disinterest (the subjective circumstance of justice) means essentially that they do not take an interest in each other's interests. Mutual disinterest has two meanings. It suggests that people are not moved by altruism, for, as Rawls argues, we do not need to think of people as benevolent for justice to be possible. It also suggests that people are not envious: they are not made better off by a worsening of another's situation. They just prefer a greater share of the social benefits to a smaller one as long as this improves their expectations towards the furtherance of their system of ends or conceptions of the good.

Having in mind the circumstances of justice thus described, with a view to choosing commonly acceptable principles of justice, we want to find the reasonable conditions one should impose on anyone's choice of principles if we want these principles to meet the above formal conditions involved in the very concept of right or in any ethical principle. (And especially if as Rawls does, we want the principles to embody some arguably widely acceptable substantive features as well.) We also want to know how persons submitted to the subjective circumstances are regarded when performing the choice of principles, that is, we need a proper description of the motivation of the parties to the agreement. All this amounts to constructing a 'fair initial situation' for the choice of principles, a situation where agreements reached at are considered fair. Rawls calls this situation the 'original position', a properly shaped starting point for the original agreement on regulative principles for the social cooperation.

The main features of this position are the 'veil of ignorance' assumption and a description of the choosers as free and equal 'moral persons'. In accordance with the circumstances of justice, the veil of ignorance is to address the issue of identity-conflict involved in social cooperation. The description of the choosers as moral persons is to address the question of the admissible systems of ends (as those aiming at the oppression of others would be ruled out), as well as suggesting some (regulative) ends worth pursuing (springing from a sense of justice of the parties).

The veil of ignorance represents a very important knowledge constraint which is to guarantee that the agreement on the principles will not be the result of bargaining over contingently acquired natural or social positions; it represents a situation of pure procedural justice.[10] It excludes the knowledge of particular facts concerning the parties in the original position, their natural or social characteristics, their knowledge of their conception of the good and their ends, their special psychological propensities, etc. The veil does not exclude, however, the knowledge of general facts concerning human nature, laws of human psychology, political affairs and economic theory, and the basis of social organization, and, of course, the knowledge of the circumstances of justice.

The question arises as to why it is necessary to impose such an unrealistic restriction. According to Rawls, the veil of ignorance can be seen as a requirement of proper reasoning towards a conception of justice. This condition would render one's reasoning regarding a conception of justice a rational deliberation satisfying certain reasonable restrictions and, accordingly, reaching some appropriate conclusions.[11] Moreover, he continues to argue that the veil should be regarded merely as a device for placing us in a certain perspective, which anyone might adopt at any time. Besides, it would insure that an unanimous choice of a conception of justice arises, for any person selected at random would choose the same principles. As a consequence, it would successfully eliminate arbitrariness in the selection of the principles whilst generating significant results.[12]

---

[10] Rawls characterizes a situation of pure procedural justice in contrast to perfect and imperfect procedural justice, in the following terms: "pure procedural justice obtains when there is no independent criterion for the right result: instead there is a correct or fair procedure such that the outcome is likewise correct or fair, whatever it is, provided that the procedure has been properly followed."(TJ:86)

[11] Cf. TJ, p.138.

[12] I quote from Rawls: "If the original position is to yield agreements that are just, the parties must be fairly situated and treated equally as moral persons. The arbitrariness of the world must be corrected for by adjusting the

(continued...)

In short, the veil of ignorance ensures that an unanimous, timeless, and unambiguous significant conception will be selected.

Still, one might argue whether there is anything to deliberate about after the exclusion of particular knowledge. This question is addressed by the description of the parties in the original position: the parties do not know the particular facts that compound their identities, but they know that they have interests they would want to further as best they could. With this in mind, as we have seen, Rawls supposes that the parties are rational in the sense familiar to social theory. To that, however, he adds two further clauses: that the parties are not motivated by envy, which is contrary to rationality for it tends to make everyone worse off, and that they are capable of a sense of justice, and, hence, of complying strictly with the agreements they enter into:

> it means that the parties can rely on each other to understand and act in accordance with whatever principles are finally agreed to.(...) the parties have a capacity for justice in a purely formal sense. (TJ:145)

What interests animate the rational parties, then? Rawls supposes that the parties are 'moral persons'. To describe the choosers as moral persons means essentially that the rational and mutually disinterested parties that are to reach an agreement are seen as being free and equally possessing some moral **powers** they want to effect. That is to say, they are capable of having ends and a conception of their good as well as having a sense of justice that makes them abide by the fair agreements they have entered to, and have a (regulative) interest in realizing these powers. As a consequence of this starting point, different systems of ends and conceptions of the good should be seen as equally worthy. Besides, though they have different conceptions of the good, the parties recognize that to realize these conceptions they need the cooperation of others so that they must abide by, and expect others to do so, the fair agreement on principles of justice that they shall eventually reach and which shall enable their cooperation. The sense of justice ensures that the agreement will be strictly complied with. Again, people so defined are understood only

---

[12](...continued)
circumstances of the initial contractual situation. Moreover, if in choosing the principles we required unanimity even when there is full information, only a few rather obvious cases could be decided. A conception of justice based on unanimity in these circumstances would indeed be weak and trivial. But once knowledge is excluded, the requirement of unanimity is not out of place and the fact that it can be satisfied is of great importance. It enables us to say of the preferred conception that it represents a genuine reconciliation of interests." (TJ:141/142)

in terms of 'capacities' and not as 'actualities', a condition which is warranted by the veil of ignorance which prevents people from catching a glimpse of the content of these 'capacities'.

The question remains what the preferences of these persons could possibly be, what would be the rational ones, that is to say the goods (and their ranking) they would prefer to have more than less of in the absence of a knowledge of the very content of their particular ends or conception of their good. These are solved by Rawls in terms of social background conditions, of liberties and opportunities, and all-purpose means, things that everyone would prefer to have more than less of regardless of his particular conception of the good, for these are things they would need to possess in order to realize their interest in furthering their moral powers. These are called the 'primary goods', and they are ranked by these persons in a certain order of priority which is, in turn, reflected in principles of justice and the way they are ordered.

In this regard, Rawls claims that primary goods solve the problem of interpersonal comparisons of well-being which has thus far had no utilitarian solution. On this account, he argues, we would not need to make reference to any particular or aggregate system of desires or preferences in order to find a suitable index for comparison. All we need is to establish what is reasonable to count as the appropriate claims in questions of justice, expressing the objective needs of free and equal moral persons, in particular, the social background conditions and the all-purpose means for any conception of the good to exist and flourish. These goods should emerge from a reflection about the needs of free and equal moral persons - capable of a conception of the good and of a sense of justice - and from a recognition of the regulative primacy over any other consideration of the interest they possess in exercising their moral powers:

> this implies a regulative desire to conform the pursuit of one's good, as well as the demands one makes on others, to public principles of justice which all can reasonably be expected to accept. (SU:165)

The primary goods account, in contrast to utilitarianism, Rawls claims, reflects an understanding of the hardship of comparing the intrinsic value of different systems of ends or ideals of happiness. A reliable agreement should, instead, concentrate on the preferences of moral persons in the original position. In particular, given their sense of justice, the citizens in a well-ordered society regulated by the principles of justice are assumed to be capable of moderating their claims in

consideration of what they have agreed upon as the legitimate claims regarding questions of justice.

So, 'the persons in the original position try to acknowledge principles which advance their system of ends as far as possible'.[13] They are moved by an interest 'of' a self, not necessarily egoistic, and of necessity, not an interest in others' interests. Moreover, in the context of the deep ignorance that surrounds their choice, they are moved by a regulative interest of realizing their **moral** personality, of being a certain kind of person, a purposeful and responsible one.

In short, the primary goods are the index of legitimate interpersonal comparisons among persons constructed as moral persons and choosing principles of justice in a properly constrained situation. The problem of trust that might arise on the way to social unity is, in this conception of justice, for the most part addressed by the constraints represented by the veil of ignorance assumption as well as by means of assuming a sense of justice as an important part of people's motivation.

3. Interpersonal Comparisons and Intrapersonal Deliberation (I): the deductive argument and the veil

3.1. The general scenario

The way the original position is constructed is to ensure that by placing the parties symmetrically, a problem of **justification** (to others) is turned into a problem of (individual) **deliberation**. Moreover, we may think that in reasoning about **public** principles, the parties will achieve also a greater **self**-understanding. To think of the parties as rational means not only that they further their ends but also that they are capable of strictly complying with the fair agreements they have entered into so that a sense of justice is made part of the parties' motivation. Furthermore, one may think of a **conception** of justice as being the mere **articulation** of a (educated) person's sense of justice, in that, in Rawls' words, everyone has within himself the whole form of a moral conception. Different as they may be, these arguments all reflect a conviction that reason at the individual level has a peculiar access to an ideal community of individuals through principles, and

---

[13] Cf. Rawls (1971), p.144.

that these principles are articulable at the level of individual rationality. Or, in our terms here, interpersonal comparisons may be solved in terms of an intrapersonal deliberation.

Now the problem of deliberation has a unique determinate solution, according to Rawls. So, we should like to understand Rawls' idea of the principles of justice as the right criterion for IC as the unique rational outcome of a peculiar process of deliberation. Given the objective and subjective circumstances of justice, the restrictions imposed by the original position regarding the knowledge conditions as well as the interests and beliefs of the parties, and, still, given some formal constraints which are associated with the very concept of right, which principles of justice would result from the rational choice of the parties? This problem, according to Rawls calls for the tools of two distinct theories, social theory and moral theory.

Social theory would yield the tools of equilibrium analysis, in the sense that it would give us the feasible solution of a standard problem of rational choice; given the options open to the individuals, it would recommend the best action to further one's interests consistent with the actions taken by others. The equilibrium framework fails, however, to provide a just or right solution on its own: it would bring about possible coordination outcomes but would be silent as to their relative justice or rightness. Moral theory, on the other hand, places constraints on the equilibrium analysis, in the sense that it would help us to select those outcomes which are just or right according to its standards. Indeed, moral theory would produce the background circumstances from which to assess the different equilibrium outcomes. Thus, the constraints which the original position is to embody come from moral theory:

> the philosophically favored interpretation of the initial situation incorporates conditions which it is thought reasonable to impose on the choice of principles (...) it is a state of affairs in which the parties are equally represented as moral persons and the outcome is not conditioned by arbitrary contingencies or the relative balance of social forces. (TJ:120)

The principles of justice would be chosen in the original position. That is, the rational parties seen as moral persons under circumstances that would be free of arbitrariness, would choose as the first principle of justice the principle of equal freedom to regulate their basic rights and duties, and as the second the principle of fair equality of opportunity and the difference principle to regulate

the division of social advantages. They would also agree to a priority rule so that the principles should be serially or lexicographically ordered: it would give precedence to the first principle over the second and to the first part of the second principle over the second part. In other words, they would agree to a conception of justice where everyone is entitled with an equal freedom and fair opportunity, and where social and economic inequalities are to everyone's advantage, or to the advantage of the less well off in the society.

We now turn to the 'deductive argument' where these principles of justice, of the legitimate IC, are derived from a deliberation procedure of a rational individual.

3.2. Ignorance and incomparability: the maxim and the primary goods index of IC

Although a number of intuitions underlie these principles, we shall concentrate on what Rawls calls the 'decisive' arguments for the two principles which aim to be entirely deductive. In addition, this will give us an opportunity to contrast the Rawlsian principles of justice with the also deductive argument for the utilitarian average principle which is based, according to Rawls, on a misunderstanding of the proper restrictions that one should impose on one's choice of principles of justice.

According to Rawls the two principles of justice may be seen as the maximin solution to the problem of social justice. Generally speaking, three features of choice situations justify the use of the maximin decision rule, namely, the impossibility of calculating probabilities, the unwillingness to take chances on a better outcome that might jeopardize a guaranteed minimum, and the rejection of certain alternatives. For Rawls these three features are present in the original position, and, taken together, they fully justify the maximin rule as the rational decision rule in that context.

Rawls argues that the maximin is not a self-evident rational rule, but looks more like a maxim or a 'rule of thumb', 'that comes into its own in special circumstances':

> Its application depends upon the qualitative structure of possible gains and losses in relation to one's conception of the good, all this against a background in which it is reasonable to discount conjectural estimates of likelihoods. (TJ:155)

In accordance, the original position has been shaped as a situation of severe ignorance in which the knowledge of likelihoods is excluded, for the parties have no basis on which to determine them:

> Not only are they unable to conjecture the likelihoods of the various possible circumstances, they cannot say much about what the possible circumstances are, much less enumerate them and foresee the outcome of each alternative available. (TJ: 155/6)

Also, the parties want to assure a minimum. In this connection,

> [t]he minimum assured by the two principles in lexical order is not one that the parties wish to jeopardize for the sake of greater economic and social advantages. (TJ: 156)

Besides, 'other conceptions of justice may lead to institutions that the parties find intolerable',[14] like restrictions on the equal liberty of the individuals.

Thus, ignorance plus the importance of the issue at stake recommend a maxim, that is, a principled choice that rejects certain compromises. In the end, the maximin ensures that the greater outcome will be selected amongst the worst possible alternatives.[15]

Now it is well known, and our study of Harsanyi's ideas seems to confirm it, that a similar knowledge restriction is imposed on individuals in the utilitarian average principle framework. So, in the reasoning leading to the average principle Rawls identifies three veils hiding particular information from the choosers, much in the same line as his veil: they do not know their places in the society and ignore their natural abilities, they have no access to the preference systems of the members in the society nor that of a deciding person, and have no particular knowledge about the structure of the societies. Considering that the average principle framework does not make

---

[14] Cf. Rawls (1971), p.156.

[15] Just as the maximin has an analogy with the rational decision rule in the choice of principles of justice in the original position, so too the parties choose analogously with risk-averse choosers. However, as before, this does not mean that the parties have a particular psychological attitude but rather that the original position is so tailored as to incorporate the three restrictions on choices that render the maximin appealing. In this way, the argument by Harsanyi to the effect that the maximin assumes risk aversion on the part of the choosers misses the point.

any special allowance for people's motivation (unlike Rawls who requires them to be choosing as moral persons), except for requiring that there be genuine and informed-preferences (Harsanyi), such people under the above knowledge restriction would choose the average principle:

> The average principle appeals to those in the initial situation once they are conceived as single rational individuals prepared to gamble on the most abstract probabilistic reasoning in all cases. (TJ:166)

The utilitarian view understands that individuals under such uncertain choice would resort to the principle of insufficient reason to estimate the likelihoods of the various circumstances they might happen to be in after the veil were lifted. By attributing equal probabilities to the occurrence of the various circumstances, they want to ensure, Rawls points out, that in the absence of any information no alternative is discarded. However, is this a sensible view? Considering the characteristics of the original agreement, Rawls argues, the answer is 'no':

> I shall assume that the parties discount likelihoods arrived at solely on the basis of this principle [the principle of insufficient reason]. This supposition is plausible in view of the fundamental importance of the original agreement and the desire to have one's decision appear responsible to one's descendants who will be affected by it. (TJ:169)

The use of probabilities in the circumstances of the original agreement on principles of justice is thoroughly ruled out by Rawls; this prohibition refers not only to the Laplacian probabilities, and objective probabilities, but also to the subjective neo-Bayesian probabilities that attempt to systematize people's intuitive estimates. These latter are entirely out of place, Rawls claims, for there are no known particular facts that could provide a ground for these judgments. Referring to the Bayesian probabilities, Rawls declares:

> Surely it is better when possible to use our intuitive knowledge and common sense hunches in a systematic and not in an irregular and unexplained manner. But none of this affects the contention that judgments of probability must have some objective basis in the known facts about society if they are to be rational grounds of decision in the special situation of the original position. (TJ:173)

At this point it is worth remarking on a very important distinctive role that the veil of ignorance is to play in the Rawlsian framework in contrast to its use in the utilitarian context. To deprive the chooser of the knowledge of particular facts is not done with the intention of denying him

particular reasons from which to choose, but rather the intention of **leaving** certain reasons as candidates for having intrinsic value, so that they would be reasons everyone could adopt. Thus the original position wants to produce the kind of circumstance conducive to **that** state of mind. In accordance, the analogy with the maximin aims to point out that there are some values that admit of no trade. So, we strive for a principle that may guarantee them.

A second problem with the average principle, Rawls contends, refers to the way expectations can be formed and known, and therefore the way the interpersonal comparisons of utility may be given sense, given the restrictions of the veil of ignorance. Rawls believes that his primary goods index is a sounder approach to the question of IC than the average utility one. Following Rawls, the utilitarian choice is thus described:

> the individual is thought to choose as if he has no aims at all which he counts as his own. He takes a chance on being any one of a number of persons complete with each individual's system of ends, abilities, and social position. (TJ:174)

A question remains however, Rawls contends:

> We may wonder then whether this expectation is a meaningful one. Since there is no scheme of preferences by which its estimates have been arrived at, it appears to lack the necessary unity. (ibidem:174)

How this expectation is arrived at depends on the individual's capacity to evaluate the other individuals' detailed circumstances, but this is not without its problems, again, given the veil restriction:

> It suffices to observe here that what we cannot do is to evaluate another person's total circumstances, his objective situation plus his character and system of ends, without any reference to the details of our conception of the good. If we are to judge these things from our standpoint at all, we must know what our plan of life is. The worth to us of the circumstances of others is not, as the constructed expectation assumes, its value to them. (ibidem:174)

The primary goods account, then, comes to address an important degree of **incomparability** at the level of the individuals' diverse conceptions of the good, which Rawls compares to 'styles of art':

> it seems pointless to try to define a measure between persons which includes the full range of final ends. The problem is similar to comparing different styles of art. There are simply many things in which human beings become engaged and find fully worthwhile depending upon their inclinations. (ibidem:174/5)

This argument by Rawls stresses the irreducible diversity of people's conceptions of the good, which he deems a fundamental difficulty for utilitarianism in so far as this latter relies upon a conception of a rational good. In this respect, it is worth reporting Serge Kolm's[16] suggestion that people may agree to a sort of shared highest order preference, a function whose arguments would be the common parameters of everyone's preference functions.

Rawls (SU, 1982) looks into this in great detail. He concludes that such a proposition fails to come to terms with the irreducible diversity of the conceptions of the good that characterizes a deeply divided society. In this latter people would fail to agree on common terms to adjudicate their conflicting claims:

> [their] final ends and aspirations are so diverse, their specific content so different, that no common basis for judgment can be found. (SU:180)

Instead of supposing shared preferences, Rawls assumes the possibility of a shared conception of **justice**, which accepts that the 'self' has a precedent **moral** structure relative to a bare satisfaction-seeking self, capable of giving priority to equal basic liberties over preference satisfaction. In this way, the primary goods account may provide a basis for interpersonal comparisons which is compatible with 'autonomy' (with respect to the precedence of a 'moral' self) as well as 'individuality' (in the form of an irreducible plurality of conceptions of the good) within the limits of a conception of justice. A workable list of these goods is arrived at from a '**political** conception of citizens as free and equal':

> It is this political conception of persons, with its account of their moral powers and higher-order interests, together with the framework of goodness as rationality and the basic facts of social life and the conditions of human growth and nurture, that provides the requisite background for specifying citizens needs and

---

[16] The reference is in Rawls (SU, 1982).

> requirements. All this enables us to arrive at a workable list of primary goods. (PRIG:255[17])

A shared conception of justice, defining the primary goods as the index of interpersonal comparisons is to be worked out not starting from the diverse conceptions of the good but from a conception of person that gives precedence to our 'moral structure'. Still, the later Rawls does not want to think of this moral structure as a transcendental structure which would still refer to a particular comprehensive moral doctrine, as in his Theory, but rather as something that acquires sense from the backdrop of a certain public culture, as in his later writings. Let us now turn to this view.

4. Interpersonal Comparisons and Intrapersonal Deliberation (II): the 'political' moral doctrine and the condition of publicity

As a matter of course, the more deeply divided the society, the less stable the social arrangement and the stronger the demands on a conception of justice. Moreover, a conception of justice inevitably excludes some conceptions of the good as altogether unfit for the social arrangement, while at the same time encourages other conceptions that appear to be more supportive of the public conception of justice and the way society is arranged.

Rawls, in his "The Priority of Right and Ideas of the Good" (1988), tries to accommodate his conception of justice in between the demands of 'neutrality' among different conceptions of the good and those of a well-defined conception of it. There he rejects the idea of full neutrality of a conception of justice, for it is not possible that it be neutral with respect to its final consequences in terms of influence and effects, but both necessary and possible that it embody neutral procedure and purposes. On the other hand, to adopt the standpoint of a moral comprehensive doctrine would show a lack of respect to the irreducible fact of pluralism which characterizes our public culture.

---

[17] PRIG stands for "The Priority of Right and Ideas of the Good" (1988).

However, the objectivity concerns of a theory of justice should address these difficult questions. In between the demands of neutrality and perfectionism, a conception of justice should be faithful to some objective values whose foundations we need to articulate. This issue is developed by Rawls with the notion of an 'overlapping consensus' among the reasonable yet irreconcilable political, philosophical and religious doctrines that one expects to co-exist in the society.

In the article "Justice as Fairness: Political not Metaphysical" (PNM,1985) Rawls provides a detailed analysis of the grounds for the objectivity of a conception of justice which must articulate the latent values of a public culture, in particular of our democratic western public tradition. As a **political** conception of justice for a democratic society, 'justice as fairness' builds upon some salient basic intuitive ideas that are embodied in the political institutions of a democratic constitutional regime and the public traditions of interpretation of these ideas:

> Justice as fairness is a political conception in part because it starts from within a certain political tradition. We hope that this political conception of justice may at least be supported by what we may call an 'overlapping consensus', that is, by a consensus that includes all the opposing philosophical and religious doctrines likely to persist and to gain adherents in a more or less just constitutional democratic society. (PNM:225/6)

The 'political' attribute is to oppose the 'metaphysical' in the sense that a conception of justice is presumed not to express the values of any moral comprehensive doctrine that makes reference to a particular conception of the good but to express in the form of regulative principles political ideals and values that underlie a democratic regime. What is more, this political conception must be the basis for an 'overlapping consensus':

> this political conception needs to be such that there is some hope of its gaining the support of an overlapping consensus, that is, a consensus in which it is affirmed by the opposing religious, philosophical and moral doctrines likely to thrive over generations in a more or less just constitutional democracy, where the criterion of justice is that conception itself. (IOC:1[18])

---

[18] IOC stands for "The Idea of an Overlapping Consensus", 1987.

So, a political conception wants to articulate ideals and values of our public culture in a way likely to be supported by the opposing views that co-exist in society.[19]

The possibility of an overlapping consensus is grounded in a restatement of the Humean circumstances of justice in terms of circumstances of 'political' justice (the social and historical conditions of democratic societies), namely, the fact of enduring pluralism that can only be overcome with the oppression of the state, and the fact of moderate scarcity, that there are 'numerous possibilities of gains from well-organized social cooperation, if only cooperation can be established on fair terms.'[20] These conditions together with the assumptions of moral psychology, 'that is, a psychology of human beings as capable of being reasonable and engaging in fair social cooperation'[21] make room for the overlapping consensus or how 'political allegiance' is generated.

In his "Kantian Constructivism in Moral Theory" (1980), Rawls produces an array of arguments for his conception of justice whose grounds are given by what he calls a Kantian notion of moral objectivity. This latter amounts to acknowledging that moral objectivity is bound to the **construction** of a 'point of view', apart from which no further moral facts are said to exist. This time, justice as fairness and the principles it yields are vindicated in terms of a different analogy from the earlier one (with rational choice theory and the maximin decision rule), namely, with Kant's moral theory.[22]

At any rate, the core question now amounts to establishing the basis of acceptance of certain principles from a recognition of certain qualities of persons, from the supposition that these persons have a common desire for agreement as well as there existing principles and notions that

---

[19] The ideal of an overlapping consensus gives primacy to the fact of pluralism which is the hallmark of a democratic constitutional regime. Therefore, a close neighbor is the notion of a *'modus vivendi'*, which Rawls rejects on the ground that such a compromise carries an undesirable instability to the social arrangement. On the other hand, the more stable the *'modus vivendi'*, he contends, the more likely that it will have acquired the features of an overlapping consensus.

[20] Cf. Rawls, IOC, p.22.

[21] Cf. Rawls, op.cit., p.22.

[22] Here, as elsewhere, Rawls stresses the fact that this is just an analogy as there are a number of differences, the main one, it seems to me, being the fact that individuals are supposed to act as if led by a pure practical reason, actually thanks to the restrictions imposed on them.

they implicitly share, which can be made explicit if necessary. To this end, Kantian constructivism selects a particular conception of the parties to the agreement, as **free** and **equal moral** persons.

The question arises how the **freedom** and **equality** of citizens as moral persons could be better safeguarded in the society. Since there is no agreement in democratic though: concerning the best institutional arrangements to that end, a Kantian conception of justice asks for **traditional** principles that would be accepted by persons seen as **free and equal moral** persons and thought of as citizens in a ongoing society over a complete life. Political philosophy should make explicit the shared notions on the basis of which principles of justice could be agreed upon:

> What justifies a conception of justice is not its being true to an order antecedent to and given to us, but its congruence with our deeper understanding of ourselves, and our realization that, given our history and the traditions embedded in our public life, it is the most reasonable doctrine for us. (KC:519)

For that matter, the freedom and equality of moral persons in a well-ordered society are represented, in the procedure of construction, as the constraints imposed on the parties as well as the way the parties are described.

As for freedom, Rawls defines the original position in terms of 'pure procedural justice', and the parties are seen as 'autonomous' in the sense of being **independent** with respect to a prior and given morality. The parties are also seen as autonomous because of the interests that animate them, the **regulative interests** in realizing their moral personality. There is an additional sense of autonomy, that of 'full autonomy', which is only effective in the normal life of citizens in a well-ordered society, where citizens effectively **act upon** the principles of justice, though its prerequisites are already present in the original position.

Thus described, the parties are subject not only to the constraints of the 'rational', that they are to act in a way conducive to the attainment of their ends, but also to those of the 'reasonable'. These latter constraints are to express the notions of reciprocity and mutuality involved in the idea of 'fair terms of cooperation'. The relation between the reasonable and the rational is as follows:

> the Reasonable presupposes and subordinates the Rational ... The Reasonable presupposes the Rational because, without conceptions of the good that move

> members of the group there is no point to social cooperation nor to notions of right and justice, even though such cooperation realizes values that go beyond what conceptions of the good specify taken alone. The Reasonable subordinates the Rational because its principles limit, and in a Kantian doctrine limit absolutely, the final ends that can be pursued. (KC:532)

In short, the parties are seen as autonomous for they have their **own** ends, **regulative** interests, and are also **reasonable.**[23] Deliberation can still take place, after final ends have been excluded from consideration, because free persons can always decide on the grounds of their regulative and effective desire to be a 'certain kind of person'. Recall that the conception of person or of a point of view is worked out as a construction, and not as rooted in any theory of human nature.[24]

As for equality, citizens are seen as equally capable of determining, understanding, and abiding by a conception of justice, as well as having an equal sense of justice. So, equality is represented in the original position by describing the parties as equal and then locating them symmetrically: 'Everyone has the same rights and powers in the procedure for reaching agreement.'(KC:550)

Now the notion of **publicity** as well as 'public reason' is of great importance in the constructive argument that is to express the basic ideas of freedom and equality. First and foremost, the condition of publicity is understood as a precondition for freedom. At the level of a **well-ordered society**, it is worked out in three ways, namely, that society should be effectively regulated by the public principles of justice; that there is an agreement concerning certain 'general beliefs' relating to human nature and social institutions among the citizens in a well-ordered society as they could be proved correct through 'publicly shared methods of inquiry and ways of reasoning thought ... appropriate'; and, finally, that there is a 'complete justification of the public conception of

---

[23] Rawls (1993b) gives the following further explanation of what is involved in the 'reasonable': "In English we know what is meant when someone says: 'Their proposal is rational, given their circumstances, but it is unreasonable all the same.' The meaning is roughly that the people referred to are pushing a hard and unfair bargain, which they know to be in their interests but which they wouldn't expect us to accept unless they knew their position is strong. **Reasonable** can also mean judicious, ready to listen to reason, where this has the sense of being willing to listen to and consider the reasons offered by others. ***Vernunftig*** can have the same meanings in German: it can have the broad sense of **reasonable** as well as the narrower sense of **rational** to mean, roughly, furthering our interests in the most effective way."(296)

[24] Indeed, Rawls develops three distinct standpoints, namely, that of the parties in the original position, that of the citizens in a well-ordered society, and the we-viewpoint, of us examining 'justice as fairness' as a basis for a conception of justice that provides a convenient comprehension of 'freedom' and 'equality'. The test for this latter is the possibility of a 'wide reflective equilibrium'.

justice...in its own terms',[25] and not just in terms of a specific conception of the good. Taken together, and given the characteristics of the conflict, the three levels of publicity can only apply 'to the public moral constitution and the fundamental terms of social cooperation',[26] concludes Rawls.

A notion of public reason is there to support our search for 'general beliefs' as well as to provide the criteria for individuals to assess the existing institutions in terms of whether they satisfy the principles of justice. Public reason is of a special kind:

> In public question, ways of reasoning and rules of evidence for reaching true general beliefs that help settle whether institutions are just should be of a kind that **everyone can recognize**. (KC:539, my emphasis)

The **publicity** condition, then, is represented in the **original position** as a condition that should preside over the agreement on principles as well as 'the companion agreement on ways of reasoning and rules for weighing evidence which govern the application of those principles.' Public reason, in this case, is limited to the shared beliefs, the procedures of science and common sense:

> The subjective circumstances of justice limit this companion agreement to the shared beliefs and the recognized procedures of science and common sense. (KC:541)

Now, an agreement on principles (as well as on ways of reasoning and rules for their application) that respects the autonomy of people turns on the possibility that the publicity condition be fully effective and public reason be in existence; in particular, the general knowledge in the original position must be equally shared or else may be proved correct to anyone with the assistance of public reason.

Here it is useful to quote the paragraph in which Rawls connects the condition of full publicity with a 'wide view of the social role of morality', in what appears to be his 'utopian' vision:

---

[25] Cf. Rawls, KC, p.537.

[26] Cf. Rawls, op.cit., p.539.

> the realization of the full publicity condition provides the social milieu within which the notion of full autonomy can be understood and within which its ideal of the person can elicit an effective desire to be that kind of person. This educative role of the moral conception defines the wide view. (KC:553)

5. Concluding Comments

The latter two sections present contrasting arguments for 'justice as fairness', the penultimate takes on a deductive argument, and therefore one that holds quite unconditionally; the last justifies the contractarian enterprise in terms of the articulation of latent values or principles embedded in western democratic culture, such as the ideas of freedom and equality.

Rawls declares, in the Political Liberalism, that his arguments in the Theory were launched from the perspective of a moral (and metaphysical) conception, a position, he claims, PL wants to redress in terms of a 'political not metaphysical' moral conception.[27] Many have wondered to what extent PL has actually amended this position, for it has appeared to some that, although Rawls attempts to enlarge the basis for his overlapping consensus, it is quite significant (and equally frustrating) that he advances **his** political philosophy as capable of articulating this basis.[28]

One may still wonder whether this is possible at all, that is, to articulate the fundamental grounds of a consensus without any reference whatsoever to one's particular place in the world. This question has been taken up by a great number of critics of Rawls' ideas on the side of the so-called communitarianism. Authors of this persuasion have called attention to the fact of the 'embeddedness of the self', that the 'self' is culturally determined and just cannot get rid of its culture to conjecture in the heights and vacuousness of an abstract person.[29] However, one might arguably ask, how can we make sense of the standpoint of these authors that are anyway trying

---

[27] Rawls versus Rawls is worth quoting: "The fact of a plurality of reasonable but incompatible comprehensive doctrines - the fact of reasonable pluralism - shows that as used in Theory, the idea of a well-ordered society of justice as fairness is unrealistic. This is because it is inconsistent with realizing its own principles under the best of foreseeable conditions. The account of the stability of a well-ordered society in part III is therefore unrealistic and must be recast...The ambiguity of Theory is now removed and justice as fairness is presented from the outset as a political conception of justice."(PL,1993:xvii)

[28] See Kukathas (1990).

[29] The 'Liberals-versus-Communitarians' debate is given a quite detailed overview in Mulhall and Swift (1992).

to articulate a general truth about our irreducible differences? It seems to me that a common ground has been denied by them, inconsistently, though. They are paying lip service to cultural differences in so far as they fail to accommodate their own view as just one amongst the many.

However, there is a more convincing way to deal with these issues which is able to conceive of 'common grounds' for social life and yet denies that individual rationality can fully articulate them. I cannot discuss this view now, although it will be addressed in the next part of this dissertation.

As for interpersonal comparability, our topic here concerns values, or, as Rawls puts it, 'these comparisons must reflect values which it makes sense to pursue'. The utilitarian cardinal project thoroughly neglects the incommensurability between distinct lives, the fact that 'happiness' is not an unambiguous unit of measurement.[30] Rawls' primary goods index wants to come to grips with incommensurability, for these goods are only social background conditions and all-purpose means, and they are to be distributed according to a certain easily identifiable order of priority, he claims. What is more, they are to reflect the fact that there are some objective values worth pursuing regardless of one's particular view. However, the problem of incommensurability is waived with the device of a supposed hierarchy, which is reflected in the design of the two principles of justice. And, this hierarchy, in turn, is supported by a certain vision of things. Can this latter conceivably come from no conception of the good at all?

In contrast, the undertaking in PL is to enlarge the basis of a consensus starting from an articulation of western democratic public culture, or 'latent values'. Still, to what extent is this articulation feasible? The terms of a possible agreement are destined to drop the substantive principles and matters of social and economic inequalities, as too much stress is put on an idea of 'public reason'. The common terms are reached at through methods and procedures of science when undisputed and common sense. To limit the agreement to what a 'public reason' might endorse has the ethical effect of trivializing the consensus in terms of freedom and a social minimum (in their more consensual forms), as well as the philosophical consequence of granting

---

[30] In this regard, Rawls states the following: "I believe that the real difficulties with utilitarianism lie elsewhere. The main point is that even if interpersonal comparisons of satisfaction can be made, these comparisons must reflect values which it makes sense to pursue. It is irrational to advance one end rather than another simply because it can be more accurately estimated. The controversy about interpersonal comparisons tends to obscure the real question, namely, whether the total (or average) happiness is to be maximized in the first place."(TJ:91)

to 'theoretical' reason the final word in the realm of practical reasoning. The former effect brings the Rawlsian view of 'order as justice' closer to a view of 'justice as order'. The latter effect, in turn, shows remarkable similarity with the utilitarian conclusions.

PART III

INTERACTING INDIVIDUALS (THE MANY)

"The incredible thing about life...is just this interaction between our actions and their results by which we constantly transcend ourselves, our talents, our gifts. (...)

This is how we lift ourselves by our own bootstraps out of the morass of our ignorance: how we throw a rope into the air and then swarm up it - if it gets any purchase, however precarious, on any little twig."[1]

"Ele não tinha ido a nenhuma parte. Só executava a invenção de se permanecer naqueles espaços do rio, de meio a meio, sempre dentro da canoa, para dela não saltar, nunca mais. A estranheza dessa verdade deu para estarrecer de todo a gente. Aquilo que não havia, acontecia."[2]

---

[1] Karl Popper, Objective Knowledge.

[2] Free translation: "He had not gone to any place. Just did the invention of remaining within those spaces of the river, from the middle to the middle, always inside the canoe, never to go away from it ever. The strangeness of that truth bewildered people. That which did not exist, did happen." João Guimarães Rosa, "A Terceira Margem do Rio" (The Third Bank of the River).

INTRODUCTION

In this Part we are going to examine Hayek's epistemological premises as well as the conditions for action in an extended order he sets up, as these are important elements of his view of social order. In relation to the previous parts of this thesis, Hayek's undertaking is clearly not reductionist in that he sees the social order not as a strict consequence of individuals' rational choices. In fact, he sees these choices as being already conditioned by evolving intersubjective multi-level rules that are not fully articulable at the level of individual consciousness. This position by Hayek concerning our limited epistemic condition also imposes limits on reason's normative reach, where rules that we follow without being fully aware of like our sense of justice cannot similarly be articulated so as to provide a firm basis for a well-ordered society. However, the exercise I undertake here concludes that, although Hayek can quite safely affirm this, he cannot as safely rule out reason's inclination for design on the grounds of the perversity of the consequences.

So, two major purposes are pursued, namely, to retrieve Hayek's inspiring insight concerning the knowledge conditions within the social world, and to so reconstruct Hayek's argument as to identify the elements within it that might favor a reconciliation with designed intervention.

The arguments for design shall be taken from within the very tradition Hayek is said to dwell in, in particular, from pieces of Karl Popper and Michael Polanyi's works. Two major results are that such a reconciliation is possible, and that it entails a revision both of traditional modes of interpreting Hayek's undertaking and the meaning of conscious action so that these might accommodate what turns out to be a synthesis or a 'third bank' of the river of social philosophy, between, that is, reason and tradition.

As for the reconciliation between Hayek's social philosophy and design, my exercise concludes rather negatively that design is not thoroughly ruled out from the Hayekian epistemological premises (actually the Hayekian actors are constructivists without success); the Hayekian epistemological premises cannot disentitle any action on the ground of its consequences, because

they entail a radical ignorance of them. As for the interpretation of Hayek's epistemological premises I have undertaken here, it stresses the cognitive nature of the social world and the fact that knowledge within this world is often uncertain, as well as the corollary fact that in trying to acquire knowledge we end up by adding to the amount of knowledge to be mastered within the social world itself. Social world is only possible because of stocks of knowledge concentrated on intersubjective spaces, such as traditions, habits, institutions, rules of every sort, which acquire in the Hayekian world a rather ontological status. However, in contrast to traditional sociology, these social forms do not dominate over the individuals determining their actions, on the contrary, the individuals perform a very important role both in apprehending intersubjective rules and adding to them. Popper and Polanyi's ideas add to those of Hayek's in conveying a more vivid image of a social world grounded on a Hayekian foundation where nevertheless there may be more active individuals. Then we come to the 'third bank of the river' of social philosophy which, though non-existent, may render us apt to dream about a world in which we are not alone since we have the company of our contemporaries, predecessors and successors and we may also be creative.

In the course of my exercise on Hayek's philosophy I have come up with an intuitive phenomenological-like approach of his conception of social order. Steven Lukes foresaw this trend and suggested to me the writings of Alfred Schutz. In fact, Schutz's rather original phenomenological approach to the social life bears a striking similarity to that of Hayek's in many respects, to a number of which I make explicit reference in footnotes. I shall not, however, undertake here a thorough reconstruction of Hayek's argument in terms of Schutz's philosophical contribution, which would be alien to my purpose here, although the points of contact are so numerous that they are worth mentioning and even worth pursuing on another occasion. Suffice it to say that the phenomenological approach to the social world undertaken by Schutz represents, along with that of Hayek's, a major challenge to behaviorist perspectives which take as given what are for the former the very starting points of our efforts to understand the social world, that is, what the conditions for individual rational action are, and how knowledge is acquired in our social world. As thinking is taken as a form of action, these questions refer to the limits of our understanding both as actors in the social world and observers thereof.

Before proceeding to the description of the following sections, I want to warn that this exercise does not intend to offer an interpretation of Hayek's work, not even an idea of what Hayek meant by his argument on the crucial role of knowledge. I realize that there are important tensions and incoherences in his work in this and other respects. I have tried, however, to offer the best Hayekian arguments in favor of the convincing premise of radical ignorance, and have attempted to separate this set of ideas from Hayek's more conservative ideas, an undertaking which I feel is possible and does not commit anyone who accepts the Hayekian premises to the Hayekian conclusions. Elements for different conclusions are extracted from works by some of Hayek's partners of paradigm, such as Karl Popper and Michael Polanyi.

This Part is organized as follows. In section 1, a picture is given of the broad landscape where the key features are transcendence, emergence and unconscious rules. It is contended that Hayek is proposing a particular view of transcendence, as the possibility, that is, that an order of radically ignorant individuals may come to emerge. This section touches upon almost the entire range of issues entailed in this picture without looking into them in depth, though. This latter task is undertaken by the following sections.

Section 2 undertakes the question of who is the subject of such an order and what are the limits to social theory which are implied by this view of the social actor. This section stresses the non-reductionist move taken by Hayek under the image of 'interacting individuals' and the suggestion he makes to the effect that intermediate formations like institutions and traditions, or multi-level rules, should be the objective of our explanatory purposes.

Section 3, in turn, outlines the problematic character of the rules themselves and the related issue of interpretation. Section 4, then, turns again to Hayek's view of individuals in order to see what their capacities are which they might count upon whenever they have to overcome problems of interpretation of the rules.

Section 5 explores the ambiguities of the Hayekian concept of order and the room given for change to come about. In section 6 it is argued that we should explore the possibility for design to interfere in the order on the basis that tradition may encompass contradictory forces of preservation and change, and among them even entropic forces. Another sense of tradition is suggested which gives consciousness an active role.

Lastly section 7 sets up the elements of a dynamic interplay between the conscious and the non-conscious parts of our actions and asserts that some indeterminacy is better than an account that leaves aside one of the following two intuitions, radical ignorance and the desire we have to overcome it.

Chapter 7

INTERACTING INDIVIDUALS

1. Transcendence as emergence[1]

We are going to examine one particular sense, within the individualist tradition, in which we might say that we can lift ourselves by our own bootstraps: we might do it without purpose. Transcendence would then be possible through **emergence** (instead of the purposive action of the individuals), as the product of human action though not of will.

To be sure, what we are to examine is the contention that if transcendence is to have any relevant meaning it must be conceived within the human circumstances of radical ignorance, as the possibility of any persisting 'order' coming about among radically ignorant men. How is this possible?

When we imply that individual purposive action is not able to generate transcendence we might think of another subject being suggested as a candidate to effect it, such as 'society', or 'history'. But the odd thing about the framework we are going to examine is that within it the individual is still the protagonist of the capacity to produce something, like social order, which is bigger than himself. However, it does not rely upon the usual division of our motives, for example interests and morals, body and soul and so on, as the basis of any possible self-transcendence.

So, if the relevant action is not purposive and the relevant subject or actor is still the individual, we are invited to think of some kind of non-conscious 'action' of his as relevant to supplement conscious and purposive actions towards an understanding of a persisting social interaction among

---

[1] In the following I shall add some comments on the text in a smaller letter whenever a clarification is in order. As for the footnotes, they will be included as presently non-explored extensions to the main text, or else as additional references to what has been said.

many individuals. Moreover, it is contended that purposive conduct itself is not a fully conscious and articulable conduct. What we deem to be purposive action is actually an action with deeper and hidden roots.

1.1 Ignorance

Under the condition of radical ignorance, that is, that there are insurmountable positional barriers that prevent anyone from having an all-encompassing perspective of his situation, an evidence of this non-conscious supplement to our actions is given, it is claimed, by the recognition by others of our actions as meaningful. If this recognition obtains, the argument goes, we may in principle think that the action that was carried out followed some common (though unarticulated) standard, shared, that is, by both the acting person and the observer.

Clarifying this issue a bit more would require us to better specify the sense of radical ignorance. What is entailed in this latter is a statement to the effect that we have in our social world (the extended order) insuperable positional barriers. We might think that the statement 'under radical ignorance any recognition should be taken as an evidence of commonality' has almost a tautological taste, for if we occupy thoroughly non-interchangeable positions and may still make sense of each other's acts, we must have something in common in some (prior-to-experience) dimension. But we should not dismiss 'coincidence' as an alternative hypothesis. This latter point is suggested by Turner (1994) who considers the whole approach of 'tradition-social practices-presuppositions-tacit knowledge' to lead to under-determination in the epistemological level. By this contention he means that it is not possible to determine exclusively by the fact that the action is somehow meaningful to others whether an action stemmed from a commonly shared social practice or tradition. Other alternatives might be equally or even more cogent, as the one he himself gives to the effect that commonality in external performance might derive from different private habits that were somehow 'trimmed' to look alike. Still, Turner points out real explanatory disadvantages involved in the concept of sharing, like the failure properly to account for transmission and acquisition of the public thing supposedly shared. Another route is taken by Hayek: this 'sharing something' is worked out as an analogy, that is, we attribute intelligibility to other's acts according to our own framework of references (which includes habits, traditions etc.), and we engage in an interaction with him in the hope that both the questions he addresses and the answers we provide are meaningful within a presumably commonly possessed system of rules. The approach is still individualistic as Turner's; selection just chooses the more suitable answer and, in this way, selection provides an ex-post confirmation of some ways (habits, traditions) of behaving, to 'rules'. So, I take Hayek to agree with the following minimum statement of 'sharing' or 'following the same rule': To take an act as belonging to a rule implies that we expect the act to commit the acting person through time within a social context, that is, that we expect certain actions to follow it, as Winch's (1958) Wittgenstein.

Now, we, the observer and the observed person, need not necessarily explicitly know the rule the action followed. Nor do we need to be able to spell out what particular instances of the action have led to our recognition of it as an action of a particular kind. Our knowledge may well have been

confined to the recognition of the action as belonging to a certain pattern that we both recognize without being required to fully articulate the knowledge we proved to possess.

> What we recognize as purposive conduct is conduct following a rule with which we are acquainted but which we need not explicitly know. Similarly, that an approach of another person is friendly or hostile, that he is playing a game or willing to sell us some commodity or intends to make love, we recognize without knowing what we recognize it from. (S, 55)

We may introduce here, as an instance of the foregoing, Polanyi's (1958) idea of the character of 'unspecifiability' involved in the mastery of some arts, in skill. Some arts are unspecifiable in detail, so transmission would be confined to example, to personal contact, and not to prescription. Two instances of related difficulties are the lack of formal guidance towards the articulation of a spatial topography from abundant access to its particulars (parts to whole), and the symmetric difficulty in specifying the subsidiary particulars that dwells in a maxim (whole to parts). These two symmetric limits to articulation reveal in the suggestion of Polanyi that the gap in between **recognition** and (articulate) **knowledge** is filled by a subsidiary awareness we develop in the activities and acts we are focally aware of. Polanyi goes even further to suggest that the act of trying to shift our awareness from its focus to its subsidiary particulars could destroy the very context that evokes the flow of the activity be it the understanding of a gesture, the playing of a piece of music, the act of speaking etc.

We might still think that conscious actions and non-conscious rules are standing for traditionally appointed cleavages of our internal motives or purposes, such as selfish and moral motives, thus keeping our actions characteristically purposive. But the idea here is rather that the non-conscious rules that we follow in fact work as a meaning-giving framework to our conscious conduct: they indicate a place to our conscious action inside a broader and unarticulated system of rules.[2] The snag, it seems, is that the knowledge assets (knowledge of various norms) which are the presuppositions of our actions, and which are situated at both the conscious and non-conscious levels, are bigger than our capacity to understand and articulate them at the conscious level.

In Hayek's words:

> there are many grounds which make it probable that, in order to be conscious, processes must be guided by supra-conscious order which cannot be the object of its own representations. (S,61)

---

[2] We might see this meaning-giving framework as still normative, for it provides norms of locating elements within an order. Yet this order is not supposedly static, for it might happen that the order be modified by a challenging element within it, as we shall see later.

A logical reason is then:

> if 'to have a meaning' is to have a place in an order which we share with other people, this order itself cannot have meaning because it cannot have a place in itself. (S,61)[3]

From this knowledge condition two immediate consequences follow: in acting, (1) we know (in the non-articulable sense - N) more than we know (in the articulable sense - A); (2) we cannot know A what we know N.[4]

These consequences are practical and theoretical, quite analogously to Kant's distinction between practical and theoretical reason. The important difference is that for Hayek practical rationality is an intersubjective and evolutionary capacity placed in an individual's mind, and theoretical reason is confined to the narrow limits of individual's consciousness. It should be also clear that the 'cannot' in the statement above stands for a **logical** impossibility.

Of course many rules may be followed consciously. The point here is that not every rule which is followed may be conscious, there will always be some rules which will be followed without being articulated for they are the very condition of the possibility of action (and, more generally, of thinking).

In particular, the closer we are to other persons the less we have to rely on unconscious rules of conduct to find our way. As far as we go through larger circles, however, thus missing the knowledge of particular details and circumstances, we have no choice but to rely on more general rules.

---

[3] Compare a congenial quotation from Polanyi (op. cit.): "They (the set of presuppositions which is our interpretative framework) are not asserted and cannot be asserted, for assertion can be made only within a framework with which we have identified ourselves for the time being; as they are themselves our ultimate framework, they are essentially inarticulable."(60)

[4] In this connection, Hayek also refers to Gilbert Ryle's distinction between 'knowing that' and 'knowing how', as expressing the above contrast. In a footnote, in his Studies...(S), he refers to Michael Polanyi's contribution on the whole issue, in the latter's Personal Knowledge, chapters on 'Skills' and 'Articulation'. Cf. Studies..., p.44.

In an extended order the model of interaction can no longer be of direct concern: "man's eyes see no further. His memory holds no more."[5] Now we have a world of unknown circumstances. In fact, within this world coordination of efforts is directed to the adaptation to the unknown (and the model of interaction should take advantage of the positional differences of the parts within a system of rules of conduct).

In society, therefore, rules are supposed to supplement our knowledge of particular facts and circumstances; they make action possible insofar as they substitute some regularity for the need for detailed knowledge of the other persons and alternative circumstances. Given our ineradicable ignorance of these facts and circumstances, we follow rules (like 'customs' or 'habits') in an attempt to mitigate, through some 'routine ways', the unpredictability of our environment. We follow them because of our limited knowledge.[6]

1.2 Rationality

However, we follow rules for another reason: simply because we **are** rule-following animals, says Hayek. This is the way our mind works or has evolved to work from innate to learnt rules, that is, largely through a classificatory scheme, a network of rules or patterns that has enabled us to give meaning to our many experiences. Mind is one such a classificatory scheme that operates, on analogy of society, as a complex system.

Thus, in understanding, we have to resort to (or look for) regularities or patterns in what we find around us. The experiences we have, whose meaning is given by rules we follow, may encompass either our future or our relation to other fellow human beings. So, a rule may help us to cope with

---

[5] Cf. Vernon, 1976, p.265.

[6] In saying that knowledge is segmented and that nobody can possibly know everything, Schutz also recognizes the existence of a 'world within common reach', which, of course, leaves out part of each individual's total world within common reach. However, "this 'common' knowledge will not have to be identical in thoroughness and detail: it may range from the pragmatically limited common-sense knowledge of the 'man in the street' to the knowledge of the expert. In certain matters, the 'well-informed citizen' occupies an intermediary position between the two. What Schutz said about zones of relevance applies here in a way: each man, in terms of his whole life situation, approaches expertness in some area or other; but he is and remains the 'man in the street'

uncertain choices, in this way connecting any coherence of our choice to something beyond our conscious purpose and thus committing us to other actions. Moreover, our relationship with other people starts from our recognition of certain patterns being used by them which render their actions meaningful to us.

If we recognize other people's action as meaningful, this does not imply that these people are actually using particular rules. What is implied, instead, is just a belief in an identification with them, a supposition that we all follow the same system of rules. We may think that this identification also constitutes the basis of imitation, in situations, that is, where in trying to arrive at a decision we just imitate what others have successfully done. Actually it should be clear at this point that the contraposition we/others is just hypothetical and the argument of making sense of what others 'do' on analogy of what we 'do' may be stated the other way round, that is, we may make sense of what we 'do' through analogy with what others 'do'. This is because imitation - an innate capacity to learn from others - provides the possibility for other rules to be learnt. Again, note, however, that this identification is worked out by Hayek as mere analogy: we suppose others to be rule-governed because we are.

An interesting contrast that may help to clarify Hayek's move here is that between analogy and symmetry. Symmetry, the imaginative capacity of ours to change places with the others, is ruled out in Hayek's approach because it would entail the denial of our epistemological condition in the social world. So, along with symmetry 'sameness' is dismissed as well, every identification, therefore, is being confined to analogy. 'Sameness' is thus just presumed and, moreover, amenable to mistakes which may ultimately provoke change.

So, we follow rules because we can only understand through patterning. If what others do somehow matches our patterns, it has a meaning for us. We may then act on the basis of this belief and see how the others will reciprocate. If their reactions are still meaningful to us we may proceed, and so on.[7] If this is so, our implicit assumption has proved to work which proposed that the others were, as we are, rule-following animals, and that they were using the relevant rule.[8]

---

in many others; finally, he may strive for additional enlightenment in a few."(Wagner,39)

[7] Schutz also remarks that social action is 'meaningful action', and analyses, accordingly, the chain of actions and reactions that takes place in social action in terms of his famous dual motivation, namely, the in-order-to

All this entails that our 'rationality', the guideline we use while acting, is itself a result of an interpersonal process: it is shaped as an abstract capacity of recognition of our own and others' actions as meaningful within an abstract framework of rules we follow, though this cannot be entirely known in the consciousness.

1.3. Emergence

Now, let us attempt to draw together this foregoing talk about rules. There are, as we have seen, two different senses of our question 'why follow rules?': because of ignorance, which is our social predicament, and because of the constitution of our mind, which is our individual virtue. They may be connected through an explanation of the **emergence** of rules, that is, of rules as an emerging effect of interaction. Thus, for Hayek, rules are the result of human action; moreover, rules have evolved as a result of human **interaction** both because they have proved beneficial to it, in overcoming ignorance, that is, and because they are the properly human device to cope with the human circumstance of radical ignorance. Cultural evolution has it that we shift from the sole command of innate rules (like our instincts) to the command of learnt rules, which is very much enabled by the interactional inclination already implicit in the instinct of imitation.

---

motives (subjective) and the because-motives (objective): "A social actor, in directing himself to another person, expects to bring about a certain action by that person. The desired and expected reaction of the other, then, is the in-order-to motive of the first actor. If the other understands this intention and responds, the in-order-to motive of the initiator becomes his because motive. While at first answering because a question was asked, however, the second actor may in turn address the first one out of a now awakened interest of his own. Having thus established an in-order-to motive for himself, he provides a because motive for the first person."(Wagner,33)

[8] So, the nature of understanding is patterning, that is, giving/recognizing meaning in the world, in other's actions. The importance of 'meaning' bears on the only possible predictability within a world whose inhabitants possess limited knowledge. It refers to something that goes beyond ostensive gesture, utterances, actions, of persons we do not directly know, causing these latter to be instances of a chain of possible sequences to which we may intelligibly react. Still, Outwhaite (1975) and Winch (1958) spell out many possible senses of meaningful action, from the rather causal Weberian account to a less determinate Wittgensteinian sense; meaningful action is, at the minimum, the action that has followed a rule that may be subjected to an external check (though it needs not actually be so).

Evolving rules could be replaced by a designing mind only at the expenses of the richness and progress of the overall order.[9] First, Hayek argues, because it is beyond any individual mind's capacity to master the immense amount of dispersed knowledge that exists in society. Secondly, and more importantly, because the ignorance of the parts within a spontaneous order is tantamount to the knowledge socially accumulated.[10]

> The first peculiarity of a spontaneous order is that by using its ordering forces (the regularity of the conduct of its members) we can achieve an order of a much more complex set of facts than we could ever achieve by deliberate arrangement, but that, while availing ourselves of this possibility of inducing an order of much greater extent than we otherwise could, we at the same time limit our power over the details of that order. (S,163)

Again, this amounts to seeing the whole that emerges from the interaction of the many parts as something not reducible to their conscious and purposive actions. In this connection, having in mind the constructivist attempt to make individual reason converge to social or collective rationality, we may see that in the Hayekian world the intellectual experiment of one person facing

---

[9] Hayek acknowledges the influence of Michael Polanyi's The Logic of Liberty in this finding of his, that there is a "difference between an order which is brought about by the direction of a central organ such as the brain, and the formation of an order determined by the regularity of the actions towards each other of the elements of a structure."(S,73) Polanyi's terms are 'monocentric' and 'polycentric' orders, the latter referring to what Hayek calls 'spontaneous order'. In another passage, the influence of Michael Oakeshott is also referred to; this time the distinction is recast in terms of the 'nomocratic' (law-governed) character of a free society and the unfree 'telocratic' (purpose-governed) social order, as Hayek is expressing his own view on the principles of a liberal social order. An interesting topic of discussion would be the correspondence between Polanyi's and Oakeshott's distinction, a correspondence Hayek seems to take for granted. However I shall not take up this question, but quote a passage that shows the kind of liberalism (and individualism) Hayek had in mind: "The conception of the common welfare or the public good of a free society can (...) never be defined as a sum of known particular results to be achieved, but only as an abstract order which as a whole is not oriented on any particular concrete ends but provides merely the best chance for any member selected at random successfully to use his knowledge for his purposes. Adopting a term of Professor Oakeshott (London), we may call such a free society a **nomocratic** (law-governed) as distinguished from an unfree **telocratic** (purpose oriented) social order."(S,163) What characterizes a spontaneous order is that it is based "on abstract rules which leave individuals free to use their own knowledge for their own purposes."(S,162)

[10] This argument by Hayek is usually followed by the Hayekian curse that any pretense of fully articulating the social knowledge thus obtained would undermine the very basis of its growth. This is, however, a *non-sequitur* for from the very Hayekian epistemological premises any such prediction is impossible, as it is impossible to fully predict the whole chain of consequences of one's actions in a spontaneous order.

an intra-personal decision problem (like a risky choice) is analogous to the interpersonal level embodied in the society **only** in the sense that both levels have to recur to abstract rules placed beyond the reach of consciousness and will.[11] Thus, as intentional attempts they will clearly fail (or only succeed by accident).

We have to consider the three components that seemingly weakens the role of will. (i) the strategic component; (ii) the evolutionary component; and (iii) the role of chance. Thus, the gap in between what we do and what we consciously know we are doing is filled by the collective dimension of our individual actions. For one thing, action is individual but its results are interactional (the strategic component); for another, even the individuality of our actions cannot be taken as a given, for it is something requiring explanation as well. In fact, our individual actions are conditional upon an interpersonal presupposition, namely, the system of rules (the evolutionary component). Moreover, that some system of individual rules of conduct prove beneficial to the whole is something that cannot be a priori known and decided: the rules individuals follow must have reached a co-ordinating effectiveness for a system of rules to persist, and this can only be checked against the circumstances prevailing (the role of chance).

2.Interacting individuals
(the anti-reductionist move)

The foregoing requires us to consider the whole of social order as a complex structure for it appears to be more than the mere aggregation of its parts,[12] and yet it results from the combined actions of the many parts. The effects of the acts of individuals aiming at particular results are

---

[11] See the preceding parts of this thesis where the social order is posed as stemming from individual rational choices. In particular, see part II.

[12] Two features of complex structures are: (i) small inputs can lead to large consequences; and (ii) slight differences in starting conditions can lead to big changes in the outcomes. These make for the unpredictability of these systems. (cf. Lewin, 1993: 11)

beyond the reach of their calculations in that when these acts of theirs combine with the acts of others a wholly unpredictable effect shall result.

The effects of individuals' actions in an extended order cannot be anticipated mainly because of the 'positional' effect such that the expectations individuals form cannot but reflect their particular position within the social world and their radical ignorance of the rest. Besides, there is also the composition effect in that unforeseen results, problems and decision situations are likely to arise as a consequence of these (otherwise motivated) combined acts, that will be constantly demanding new answers, thus constituting further circumstances for individual actions.

Thus, in one sense this whole is constituted by the actions of the individuals but in another sense it 'constitutes', 'shapes' and restrains the actions of the parts in the manner of a quite autonomous structure. We have here the intersubjective space playing a rather important role, by affecting the very parameters of individual actions. What is this 'intersubjective space' like?

One claim by Hayek regarding the intersubjective space is rather a disclaimer: it is not meant to be any collective social agency as in a 'collectivist' approach, for he denies that social entities can be any more real than individuals. It seems that in saying this, he is envisaging all realistic or essentialistic approaches to social interaction, either individualistic or collectivistic. He then declares that he sticks to individualism not as in any sense a more realistic hypothesis than collectivism but as a sort of 'nominalistic' approach to social life.[13]

The adoption of philosophical nominalism is justified by Hayek in contrast to essentialism: individualism is a necessary **hypothesis** towards the understanding of social phenomena which is the 'understanding of **individual actions directed toward other people and guided by their**

---

[13] Schutz also stresses the essential character of intersubjectivity as a 'fundamental ontological category of human existence', a 'precondition of all human experience in the life world'. However, his approach, as I presume Hayek does too, differs from traditional sociology in that "the collective elements in human orientations neither eliminate individual spontaneity and volition nor even prevent idiosyncratic interpretations of cultural typifications and definitions." (Wagner,47/48)

expected behavior' (my emphasis). That is, towards the understanding of social life we should suppose 'men whose whole nature and character is determined by their existence in society.'[14] In this way, Hayek's individualism assumes individualism as an hypothesis and the individuals as social creatures.

Thus, towards our understanding of social life we cannot directly comprehend the social whole as collectivists want us to; nor can we take individuals as isolated and self-contained realities or essentials (as in the essentialistic individualist tradition). To understand individuals we need society, to understand society we need individuals. We want, then, some intermediate category as individuals in society, or individuals as **interacting** individuals.[15] This is labeled the 'true individualism'.

What is at stake, it seems to me, is Hayek's rejection of any **reductionist** approach to social phenomena either resolving social interaction as stemming from any reasonable individual's behavior or explicating individual's behavior from a comprehensive understanding of the social whole. Every understanding of social phenomena must be indirect and must take as its minimal category this intricate object constituted by individual and social interaction, a category not reducible to its constitutive parts.

In this connection, Hayek reads Adam Smith not as a defender of any natural harmony of individuals' interests but as stressing something else, namely, the role of **evolving institutions** (an intermediate category) in bringing about a reconciliation of these interests. Another instance of this 'nominalist' argument is found in Hayek's interpretation of Mandeville's work. He considers Mandeville's greatest contribution to reside in the insight that any 'rationality' that may be found in human action has its source in the **restraints** imposed on men by **institutions and traditions** of society. The alpha and omega of Mandeville's predicaments are, then, that

---

[14] Cf. IEO, p.51.

[15] Schutz pointed out, in the same line as Weber's sociology, that the relevant social phenomena are social actions. I quote from Wagner's comment on Schutz's individualistic perspective, very similar to that of Hayek's: "Schutz, of course, proceeded from a preliminary individualistic perspective to the direct analysis of social relationships. Social interaction involves the social action of at least two people who orient themselves upon each other. And living in the world of everyday life, in general, means living in an interactional involvement with many persons, being entangled in complex networks of social relationships." (Wagner, 30)

> we do not know why we do what we do, and that the consequences of our decisions are often different from what we imagine them to be. (NS, 250)
>
> (...)
>
> While he [Mandeville] still seems most concerned to show that it is merely pride (or 'self-liking') which determines men's actions, he becomes in fact much more interested in the origin of the rules of conduct which pride makes men obey but whose origin and rationale they do not understand. After he has convinced himself that the reasons for which men observe rules are very different from the reasons which made these rules prevail, he gets increasingly intrigued about the origin of these rules whose significance for the orderly process of society is quite unconnected with the motives which make individual men obey them. (NS, 257)

Rules, we may conjecture, are there to help individuals to achieve their particular aims and help order to be preserved or enhanced thereby. That is, they are supposed to enhance local predictability and global preservation at once.

In "Between Instinct and Reason", Hayek argues that institutions and traditions, learnt rules in short, restrain our instincts or innate rules. In focussing more precisely on two major driving forces of human action - altruism and aggressiveness - Hayek makes quite clear the interplay of 'inner motivation' and 'restraints'. He seems to suggest, then, that restraints would help us in shaping both the particularity and universality of our actions. For in restraining our innate altruism learnt rules would prevent us from losing the knowledge we would be more competent to provide in our particular position within the world as well as losing the independence of our aims, the lack of which would make us the prey of particular interests of others. Note that self-interest loses its appeal as an explanatory category, and becomes almost senseless as an independent explanation of individual action. As for the universality of our actions, restraints on our innate aggressiveness would place us under general rules equally applied to everyone.[16],[17] Let us turn now to Hayek's critique to what he calls the 'false individualism'.

---

[16] It is worth-noting that for Hayek we live in two worlds at once: innate and learnt. It will be the object of specific and further consideration the possibility that we also live in a third world that would render us less passive. In Hayek's sense we may think of an economy of instincts, of some instincts, that is, as helping us to displace others, as has imitation. Imitation is thought of as a capacity to learn from others. Again, we may conjecture that this capacity is less passive than it may look, as suggested by Polanyi. See the last section.

This false individualism (of Rousseau and the physiocrats, of Cartesian origin) belongs to the constructivistic approach and, contrastingly with Hayek's account, starts from a presumption of Reason, with a capital R, of the individuals, a reason which is

> always fully and equally available to all humans and that everything which man achieves is the direct result of, and therefore subject to, the control of individual reason. (NS, 8)

It is a 'false' individualism because it leads to 'practical collectivism'. To this approach Hayek opposes his presumption of **ignorance** on the part of the individuals:

> It is the contention that, by tracing the combined effects of individual actions, we discover that many of the institutions on which human achievements rest have arisen and are functioning without a designing mind;... and that the spontaneous collaboration of free men often creates things which are greater than their individual minds can ever fully comprehend. (ibidem:51)

This latter contrast gives Hayek the opportunity to vindicate his rejection of a rationalistic approach to social interaction, as his initial presumption concerns precisely the limited capacity of individual reason to account for interaction within an extended order, either practically or theoretically.

Some of the major short-comings of individual rationality[18] seem to be, then, (1) the discontinuity between what the individual wants and what he really achieves; (2) the poor explanatory power of

---

[17] Since we dislike these restraints, Hayek wonders why they came about. His answer is that rather than being selected by our will, these rules have selected us, as the enabling survival devices they have turned out to be. This kind of contention has been identified as Hayek's 'dogmatic evolutionism' by Espada (1996) who rejects what I call the 'perfectibility' aspects of Hayek's account of evolution. I want to distinguish here Hayek's inspiring 'evolutionary approach' from his more disputable 'evolutionist' approach. This distinction is worked out in later sections.

[18] As for the limited role of man's rationality, the following passage picks out some of Schutz's insights: "Insofar as action is based on conscious planning, it has frequently been called rational. (...) In view of the existing ambiguities, [Schutz] considered Weber's concept of rational action an ideal not attainable in everyday-life conduct. Of course, he did not deny that men make rational choices in terms of their relevant knowledge on hand. Yet, he preferred to call everyday action characterized by such choices reasonable rather than rational, making allowances for the unavoidable shortcomings of practical knowledge. No individual,

psychological assumptions in the understanding both of the emergence and prevalence of any overall order; (3) the autonomy of the whole, in that the whole is ever changing as a non-anticipated aggregate consequence of the actions of interacting individuals (though its inarticulability prevents it from being a privileged agency).

Hence, towards the understanding of the social world we should cast light on the places where action is significantly taking place; we should then highlight the plausible role of intermediate formations and the scarcely significant influence of a conscious subject, either individual or collective, on the process which generate that world.

Accordingly, Hayek stresses the role of **tradition** as an intermediate category between instinct and reason, in connection with his idea of **mind** as a capacity to restrain instincts with the aid of learnt rules. Mind, for him, is nonetheless a capacity for generating testable knowledge and interpretation about the world. Individual reason, however, cannot test mind because its effects are on the group and over cultural evolution.

3.Good and bad presumptions: *Verstehen* and Equilibrium

Institutions have arisen from action, which means that the combination of individual actions has given rise to some restraints to actions over time, although in a thoroughly undesigned way. If institutions, and more generally rules are in turn the condition of the possibility of social life, that is, if they have evolved to overcome inherent limitations of knowledge available to man that otherwise would have rendered social life impossible, social theory should attempt to cast light on these institutions and practices, on the restraints, that is, rather than on the psychology or the rationality of the individuals.

Relevant questions are, then: how may ignorance have worked for social life? what kind of ignorance is it? It becomes more and more irrelevant whether the individuals are moved by pride or self-love, or by rationality and insight, and more and more relevant to understand how some restraints have been built up which enable the orderliness of our societies to arise given our limited

---

probably, is ever to have knowledge of all actually and potentially relevant factors in those situations in which he 'works' toward the realization of his plans.' (Wagner,p.26\27)

knowledge. Rather than trying to grasp human nature and from this knowledge to deduce human society, we should insert individuals in their society - as interacting short-sighted individuals, groping their way, trying to guess what the others' actions will be, what the results of their acts will be, acting from rules not fully understandable through rationality - and try to grasp the constraints that have appeared as necessary to contain their ignorance within manageable limits. The initial purpose of the individual acts will practically disappear in contrast to the effects they may promote in the whole, hence their intelligibility to a distant fellow actor (or to a more sophisticated interpreter, like the theoretician) is increasingly being given by the meaning of one's action within a system of the general rules we follow.

So, let us take the question of how we can make sense of ignorance. Under ignorance we may be guided by some common rules which so far have proved to function to the benefit of the whole. The market provides an example of the preceding, for it is a mechanism, as any other institution, that is supposed to make the dispersed information possessed by many individuals available to all in the form of positive or negative incentives.

But how does it operate, this ignorance which turns us into potential consumers of rules? We can find implicit in Hayek's view the idea that our ignorance may work both actively and passively for the overall order. Taking passive ignorance first, this is an ignorance we cultivate without being aware of it, actually as a paradoxical effect of our being steeped in a particular, positional knowledge. This ignorance is tantamount to the distance we create between what we particularly know and what the others taken together know; it is analogous, in the division of labor, to the ignorance of the expert in some very specialized thing. That is, we increase the distance between us and we improve the map at the same time. So, in being ignorant in this passive sense we actually contribute to the growth of accumulated knowledge (or the complexity of our society).

We may also think of ourselves as profiting from our own ignorance in a rather active sense: when we let ourselves be guided by some practical knowledge embodied in a rule instead of a more demanding theoretical knowledge for which would need to make stressful and impossible

calculations of causes and consequences, we can possibly know more (N) than our knowledge possibilities (A). Again, not because we consciously know what we are doing but because our actions will be more knowledge-intensive, actions in reference to the others and to the whole.

In this active sense, ignorance is really a recommendation in line with Hayek's curse that we should not pretend to know (A) what we know (N), otherwise we would undermine the spontaneous order on our way to shifting from (inarticulable) rules to reason. Since it is helpless to strive on the basis of our predictive powers alone, we should consciously give them up and follow the rules. Even granted that an awkward 'active ignorance' is possible, which I doubt, an additional doubt remains as to what 'rules' we should follow. This is not altogether unproblematic.

Hayek is well aware that these 'actions in reference to others' actions and to the whole' involve a great deal of interpretation. Are these references, or rules, some sort of price system, a general objective signal? This is absolutely not Hayek's idea. Though the market image might suggest that there is a mechanism that would lead to a perfect coordination among the actions of the many individuals, Hayek is very much a critic of this view of the role of the market mechanism. For him, the market is not the place where division of labor is actualized, it is more than this; it is, more significantly, a place where the division of knowledge among economic agents is expressed, meaning that different knowledge positions (knowledge of the different alternative possibilities) are confronted, different capacities to predict and take advantage of the particular position (or information) each agent possesses. The market is a procedure of discovery.

In this sense, Hayek argues in Individualism and Economic Order that traditional equilibrium analysis begs the question when it takes for granted the knowledge people possess. The really relevant question for an understanding of economic coordination is precisely **how** people come to acquire the knowledge they possess. If this latter is really the relevant question we should try to grasp the role of institutions in affecting these bits of knowledge that each one of us possesses, thus enabling actors to alter their plans, and theoreticians to make sense of change. For change, after all, is what economic system is all about. So, in a sense, it is essential that expectations be disappointed,

either because this reflects inevitable differences in the knowledge of the agents (again, not only of current and past prices, that is, not only differential information but also of alternative possibilities to the plans the agent has designed) or because this situation enables change and progress of the whole to take place.

Here is an occasion to comment on what is to my mind a misleading interpretation of the role of market in Hayek's thought. I think that we may find in his work, as emphasized in this section, many grounds to support the idea that the market is an institution concerned with change and not a static mechanism of matching different expectations formed under ideal knowledge conditions. This latter view is implied in Vernon (1976) and Crowley (1987) when they claim that Hayek unfairly models political society on the image of a market which is characterized as a mechanism of exchange of essentially complementary and commensurable goods. A more inviting understanding of Hayek's view of the market (more in line with his epistemic premises, I believe) seems to be that he displays a very distinctive view of the market to the effect that it is concerned with a system of rules operating behind the backs of differently positioned individuals. What the market metaphor suggests to the understanding of a political society is, on the contrary, the contention that nothing is perfectly commensurable (not even goods). Actually, the market is modeled on the analogy of politics; it is the place of confrontation of different and, to a considerable extent, conflicting and non-perfectly commensurable expectations.

The difficulty of unambiguously following a rule has its roots, it seems, in the difference that there is between **symmetry** and **analogy**. We may say that according to Hayek actors are analogously positioned but not symmetrically. If this were the case we would collapse into a form of essentialism (that 'individuals' are self-contained realities or essentials) or false individualism (that starts from a presumption of Reason with a capital R) that he had rejected from the outset.

In any case, interpretation comes inevitably to the centre of our concerns for we do not have any objective and externally given mechanism of signaling that is able to get rid of the subjective evaluation of each agent concerning the interaction.[19] We cannot help entertaining expectations as

[19] In this regard, Caldwell (1994), drawing on Hayek, remarks that "(...) Hayek's larger project was to determine how the actions of agents whose subjective beliefs differ could ever become coordinated. Because of its focus on equilibrium end-states, equilibrium theory was rejected as a tool for furthering our understanding of the coordination problem. (...) Equilibrium theory assumes that the coordination problem has already been solved."(311) But then he concludes that: "If one focuses instead on how our interpretations create and change the world, then the whole question of interpersonal coordination may become problematical."(311) I do not quite agree with this second contention if by 'problematical' Caldwell means that Hayek did not account for the problem of interpretation, as shall be clear from the following parts of this dissertation, especially section 4. I do believe that the way Hayek addresses the problem of interpretation also suggests a view of 'coordination' different from that implied in standard general equilibrium analysis.

to what others know or believe, since we cannot ignore - it is indeed daily before us - that the achievement of our aims depends to a great extent precisely on the expectations entertained by others.

In implicitly acknowledging a place for interpretation, Hayek is pointing to the crucial question that the genuine problem of social life is not the amount of **information** individuals possess. Again, what is really envisaged by Hayek is rather the **capacity of knowing**. And it is not so much our limited computational capabilities that are at stake, though they play a part (as the epiphenomenon):

> The number of separate variables which in any particular phenomenon will determine the result of a given change will as a rule be far too large for any human mind to master and manipulate them effectively. (CR,42)

More important is the fact that knowledge, the stuff out of which our social world is made, is itself a complex structure with its own domain of autonomy, something that moves away from us, at the very moment we try to catch it.

Though this is not explicitly stated by Hayek, this interpretation is compatible with many relevant parts of his work and it refers to the consideration by Hayek that the social world is entirely built up of 'theories', and that any of these theories affects it unpredictably. In this way, the path to design is rather severed.

This latter consideration is bound to limit not only the ordinary theories entertained by social actors but also any social theory. So, anyone attempting to grasp this world would have to be aware that among the huge number of separate variables the theoretician is looking at he should include his own efforts. As no one single mind wields the power to gather all the required knowledge (a complex structure itself) to run or understand the social intercourse, so every single mind cannot but work on the modest presumption of ignorance of what others think and will do and, therefore, what the consequences of his own actions will be.[20]

---

[20] Yet, why add to the 'impossibility' a recommendation of **active** ignorance? It does not seem to follow

As far as social theory is concerned, this presumption is translated into the form of an explanation of the 'principle', which is only a negative form of knowledge:

> (...) While we can explain the principle on which certain phenomena are produced and can from this knowledge exclude the possibility of certain results, e.g., of certain events occurring together, our knowledge will in a sense be only negative, i.e., it will merely enable us to preclude certain results but not enable us to narrow the range of possibilities sufficiently so that only one remains. (ibidem, 42)

The way to design is blocked because the way to full prediction is similarly blocked, says Hayek, in a statement that ties our epistemic condition to limited normative prospects. Social interaction, individual actions and theoretical undertakings all remain bounded to *Verstehen*, to intelligibility, to make sense from within, for it is impossible for any part to postulate an external point of view.[21] Just as people base their actions on the perceptions they have of the meaning of other people's action, so will the social theorist take these as a datum and he cannot help adding his own interpretation of people's actions and beliefs when trying to re-construct their reasoning.[22] In

---

logically from the 'impossibility' argument and actually renders it rather awkward. The only justification I can see is that there is an implicit normative assumption which does not follow from the epistemological argument at all. In fact, from this latter we may deduce an intricate relationship between subjectivity and objectivity instead of an unequivocal convergence to an objective, and for that matter disastrous, equilibrium.

[21] The specific problem to which Hayek refers is that of the role of perception of the meaning of other people's action in the scientific explanation of the interaction of men. To quote, "The problem which arises here is known in the discussion of the methodology of the social sciences as that of *Verstehen* (understanding). We have seen that this understanding of the meaning of actions is of the same kind as the understanding of communications (i.e., of action intended to be understood). It includes what the eighteenth-century authors described as sympathy and what has more recently been discussed under the heading of 'empathy' (*Einfhhlung*). Since we shall be concerned chiefly with the use of these perceptions as data for the theoretical social sciences, we shall concentrate on what is sometimes called rational understanding (or rational reconstruction), that is, on the instances where we recognize that the persons in whose actions we are interested base their decisions on the meaning of what they perceive. The theoretical social sciences do not treat all of a person's actions as an unspecifiable and unexplainable whole but, in their efforts to account for the unintended consequences of individual actions, endeavor to reconstruct the individual's reasoning from the data which to him are provided by the recognition of the actions of others as meaningful wholes."(S,59)

[22] Schutz has a well known formula to state the case here: 'the world is a typified world', as for him ideal types pervade our world both as observers and objects of observation. To quote from a comment on Schutz by Walsh: "[t]he use of ideal types does not, then, enter at the stage when we pass from prescientific to scientific observation. It enters rather when we pass from direct to indirect social experience."(ibidem, p.xxviii)

understanding, the social theorist is actually judging people's judgments with respect to the meaning of other people's actions. And the basis for this knowledge of others is provided by rules. Even theoretical knowledge is constrained by the practical knowledge of some common rules, supposedly shared by interpreters, that is, the acting individual and the theoretician:

> We shall indicate this limitation by speaking of **intelligibility** and of **comprehending the meaning** of human action rather than of understanding. (S,59)

As far as both ordinary and sophisticated observers are concerned, every viewpoint is internal, that is, to this universe of commonly shared rules where every position is to be meaningful. Yet every evaluation is an interpretation subject to error, for intelligibility is certainly a matter of degree. The suggestion Hayek offers as a practical guideline here is that we should not dismiss our presumption of shared rules if it can be seen to work.

But, take care: the notion of equilibrium in the traditional analysis is a bad presumption, for the objectivity it entails is both misleading and undesirable. I do not have space here to explore the logical problems Hayek identifies in the notion of equilibrium[23] and which I have just hinted at, but the question concerning its desirability is internal to our scope. The reason is that the richness of the whole, something highly desirable in the Hayekian world, depends on asymmetries, different subjective evaluations of the individuals, and different circumstances of choice for each individual, all of them at odds with traditional equilibrium conditions. Disequilibrium is practically the motto of the spontaneous order.

In any case, Hayek is not claiming the end of objectivity when he criticizes the equilibrium concept, for objectivity is not altogether dismissed but just reinterpreted. In my view, Hayek presumes objectivity to be a matter of degree, being more meaningful as the boundary conditions of the very subjective assessments of the individuals. Those conditions are subject to change, but, in contrast to

---

[23] See IEO, specially p. 55.

these latter, they evolve slowly in time. Objectivity is therefore in the time-dimension, not alien to (slow) change and, in the space-dimension, quite intersubjective:

> We can judge and modify all our views and beliefs only within a framework of opinions and values which, though they will gradually change, are for us a given result of that evolution. (S, 75)

4. Individuals: Experience plus Constraints

(limiting subjectivity on the way to Verstehen)

Let us pause to consider more closely what this approach assumes the individuals to be. My hypothesis is that the individual is an indefinite extension of knowledge surrounded by the deep seas of what he ignores. More to the point, the individual is a locus of experience and constraints, a complex structure himself. He is prey to the circumstances and particular facts and it is through him that rules are experimented and tested, yet the result of his efforts is not alien to him.

The individual experiences the circumstances and the particular facts and then experiments and tests the rules, each experience being quite irreducible to any other. So, the individual is defined with reference to the whole (the system of general rules) as well as with reference to his idiosyncratic experience. It is his experience of the circumstances which commands a combination of rules to be applied and a further testing of the suitability of these rules.

On the other hand, though, the sense of an individual's experiences is already given to him by some set of largely untestable rules. So, we have two levels of rules operating through the individuals, namely, testable and largely untestable rules. Although even these latter can be tested in the long run, not all rules can be simultaneously tested because some of them are the very premises of individual experiences (but contrast Popper's view in the last section). In this connection, Hayek evokes the familiar feeling of dizziness that comes about every time we realize that the parameters of any one of our acts are missing and we experience the fear of having the ground taken from

under our feet. He takes this as an evidence that we act to a large extent on an unarticulated presumption that our action has a place within a more comprehensive order.[24]

From the preceding we can see a fundamental tension in Hayek's image of the 'interacting' individual, as the locus of both freedom and constraints that he conceives the individual to be. In short, one can say that Hayek emphasizes the role played by constraints, as in his idea of freedom confined to 'freedom to err' under these constraints, that is, freedom in a 'negative' sense. However, he also sees a quite innovative function of this freedom, in a more 'positive' sense, as it takes the form of an 'additive' capacity of the individuals when making their way in the world. Now let us try to understand the conditions of individual action and try to examine the interplay between these paradoxical elements.

Some initial driving forces, such as hunger (which is one of Hayek's examples), plus external circumstances establish the situation requiring action from the individual. He looks, then, for the rules applicable (a combination of a number of abstract rules), and acts accordingly to satisfy his desire. In doing this the individual is simultaneously making use of some explanatory knowledge, of causes and effects, that is, concerning the means available and the immediate ends he has in mind, and some normative knowledge, referring to the more abstract and unarticulated rules constraining his choices. Among these latter, there are more general rules that shape the situation as the decision problem it looks like, rules, that is, that are already at work when the individual begins to search for rules applicable to his situation. This description gives us the opportunity to examine a little further what is the 'real' action and how complex the actor may appear.[25]

---

[24] According to Schutz, in Wagner's words, "all human experiences are experiences in and of their life-world: they constitute it, they are oriented toward it, they are tested in it." (...) "The life-world ... is the whole sphere of everyday experiences, orientations, and actions through which individuals pursue their interests and affairs by manipulating objects, dealing with people, conceiving plans, and carrying them out."(pp.14/15) Man operates in this world with the help of a 'natural attitude' (pragmatic, utilitarian and realistic), 'biographical factors', and of a 'stock of knowledge on hand' (routine character). According to phenomenological theory, each individual constructs his own 'world'. But he does this with the help of building blocks and methods offered to him by others: the life-world is a social world which, in turn, is prestructured for the individual."(ibidem, pp 14/15)

[25] Even the aims of our actions do not appear either as clearly given before us or as something revealing in

4.1 Explanatory and Normative Knowledge

We have then two different 'actions' taking place inasmuch as two different sets of knowledge are required. We may refer to them as the explanatory or 'subjective' knowledge and the normative or 'objective' knowledge.[26] The explanatory knowledge is a 'subjective' knowledge in the strict sense that it is a knowledge with a knowing subject, one that requires a subject searching for means to achieve its pre-established aims, taking advantage of his concrete knowledge. It is subjective also in the sense of a knowledge of causes and their known effects, its efficiency being circumscribed to the sphere of the particular facts that regard the individuals. So:

> the knowledge and beliefs of different people, while possessing that common structure which makes communication possible, will yet be different and often conflicting in many respects (...) the concrete knowledge which guides the action of any group of people never exists as a consistent and coherent body. It only exists in the dispersed, incomplete, and inconsistent form in which it appears in many individual minds, and this dispersion and imperfection of all knowledge is one of the basic facts from which the social sciences have to start. (CR, 30)

The normative knowledge, on the other hand, is a sort of objective knowledge, a knowledge without a knowing subject in the sense that it goes beyond the subject's sphere. We may examine this surplus, so to speak, with the help of two dimensions, namely 'production' and 'consumption'. So, this objective knowledge is said to be a knowledge without a knowing subject because, though

---

terms of an understanding of social phenomena. It is true that our actions start from some more or less immediate aims, as purposive actions, that is. But the relevance of this initial impulse on the overall chain of consequences it sets in motion quickly fades away. Besides, the ultimate ends of the individuals are quite unknown to them, in the uncertain world we all live in. They are beyond their decision sphere for they are susceptible to variation according to the circumstances and new possibilities offered by the ever growing amount of knowledge that society possesses. All the individual can autonomously do is to procure himself the 'means' - that is, the immediate ends - he thinks will help him to satisfy those socially given ultimate ends, whatever they may be in the future. However, even some immediate ends take a social form: money, in an advanced society, actually one of its institutions, will function as an immediate end that works as a 'generalized means' to ultimate ends. So, the adage goes, each individual searching for a position to achieve his immediate ends (socially given as this 'generalized means' - money (a statement about which the adage has so far been silent)) will derive widespread desirable and unforeseen consequences. And in order to accomplish it, each individual ought to be free to err in his search which no one could do in his place.

[26] I am forcing here an approximation between Hayek's explanatory/normative opposition and Popper's subjective/objective opposition, thanks to which I may 'free-ride' Popper's hopes. See the last section.

resulting from the subject's endeavors to causally know something, it exceeds the subject's efforts. On the consumption side, we have to think of it as something that is already conditioning individual's explanatory efforts to address some choice situation; for it frames the hidden premises of his choice, actually constraining and reducing the scope of his free choice.

In other words, the objectivity of this normative knowledge comes from the role it plays in meaningfully delimiting the sphere of individual action (the consumption side), and from the fact that, though resulting from individual action, it is quite autonomous with regard to the initial stimulus the subjective sphere exerts on it (the production side).

We are now going to examine the delimitative function of normative knowledge, its consumption side. I shall leave the question of its autonomy to a further treatment in the next three parts, although we shall anticipate in this section the extent to which normative knowledge is constructed by individuals' expectations.

What do we mean when we say that the objective knowledge delimits the sphere of individual action? We mean that it makes us perceive some situation as we perceive it, that it gives form even to our sensory perception, and that, more generally, our sense perceptions, our actions and our perceptions of other's actions are governed by its norms. This role is largely negative:

> (Rules) will often merely determine or limit the range of possibilities within which the choice is made consciously. By eliminating certain kinds of actions altogether, and providing certain routine ways of achieving the object, they merely restrict the alternatives on which a conscious choice is required. (...) Thus even decisions carefully considered will in part be determined by rules of which the acting person is not aware. (S,57)

However negative the role of rules may be, when it comes to **internal** constraints placed in the very sensorial capacity of the individuals, this negative aspect of the rules becomes almost positive. (I wonder whether, in Hayek's view, any positive 'activity' of an action would have a positive value whatsoever, for the 'positive' aspect of an action seems to lie in a thoroughly irrelevant initial

stimulus to it. I return to this point later in this section.) Anyway, let us turn to Hayek's idea of rules delimiting the universe of our perceptions and actions.

1. Sense-perceptions (SP):[27]

Recall that we are rule-following animals; we perceive and understand the world and ourselves by resorting to the classificatory scheme our mind provides us. And thanks to it we can make sense of what we see, hear and, more generally, feel. Abstract principles or patterns superimposed on our sense perceptions and instincts provide a more sophisticated classificatory scheme than would be provided by our bare instincts, and specify "what we are to regard as objects or events of the same kind or different kinds", that is, they relate external objects to an order. Hence, our **sensorial** experience of the world, in general, turns on our thoroughly mastering a highly abstract scheme of patterns and regularities, a system of rules, in short. But, as we have already seen, this capacity of ours cannot entirely appear at the conscious level. But we can still know N it.

2. Sprachgefhhl (S):

Beside this sense-sphere, we may also recognize that we 'buy' some rules by learning a language, without being aware of it, though. So, when a child learns a language she learns more than she is able to articulate in terms of the rules (*Sprachgefhhl*) she now gets a practical knowledge of. Moreover, by learning a language we learn more than the language itself in the sense that we acquire through it a picture of the world in which our own and others' actions become meaningful.[28]

---

[27] The mark that hereinafter follows each set of rules, namely, SP, S, R, and T, are later to be used in a diagram in subsection 4.2.

[28] This passage is reminiscent of Hayek's relative Ludwig Wittgenstein, in particular of the latter's later work, Philosophical Investigations. Wittgenstein is only mentioned once in Hayek's Studies..., but most significantly, in the context of *Sprachgefhhl* and *Rechtsgefhhl*, as these are defined as capacities of ours "to follow rules which we do not know in the sense that we can state them"(S,45). The reference to Wittgenstein is in the footnote where Hayek quotes from the Philosophical Investigations the following extract: '"knowing" it only means: being able to describe it.'

> As we learn as children to use our language according to rules which we do not explicitly know, so we learn with language not only to act according to rules of language, but according to many other rules of interpreting the world and of acting appropriately, rules which will guide us though we have never explicitly formulated them. (S,87)

3. Rechtsgefhhl (R):

Analogously to our judgments of meaning, our judgments of the justice or injustice of our own or other people's actions (*Rechtsgefhhl*) are also based on our possession of some abstract and unarticulated rules; they are somehow to help us coordinate our actions to others' and to the whole according to how far these actions fit into the abstract system of the rules of conduct.

Thus we acquire a practical knowledge of grammar and other practical rules as we acquire a sense of justice, in a way, that is, which we are unable to specify and fully articulate. Both capacities, says Hayek, must be based on ourselves possessing and being governed in our perceptions and actions by a highly abstract system of rules, inhabitants of the world of normative/objective knowledge.

4. Perception of others' actions (T):

Another important instance of our actions and perceptions being rule-guided is that rules guide our perception of other people's action, the relevant world in which we act where not only are we guided by rules but we also perceive other people's actions as rule guided too. It is worth noting that here we have a different class of rules than *Sprachgefhhl*, a class which is located in a lower level of non-consciousness, because these rules are usually represented as 'customs' and 'habits', and as providing 'routine ways' of dealing with situations.

So, we are always looking for patterns and regularities in other people's conduct, and we may say that what they do has a meaning for us whenever their actions seem to conform to a recognizable rule which we judge applicable to the situation:

> we take it for granted that other men treat various things as alike or unlike just as we do, although no objective test, no knowledge of the relations of these things to other parts of the external world justifies this. Our procedure is based on the

> experience that other people as a rule (though not always - e.g. not if they are colorblind or mad) classify their sense impressions as we do. (CR, 26)

In this connection, what is important is that this common 'acquaintance' happens, in the sense that the interpreter and the people involved in a particular situation recognize it as one of a certain kind. It is not required that they be able to articulate every particular that makes the situation an instance of a more general class. The point by Hayek is that more often than not this latter knowledge does not obtain, as we have already hinted at. This is the kind of problem better illustrated by physiognomic perception, whereby we understand a gesture or a facial expression as expressing something else. But it can be extended, Hayek contends, to wider domains insofar as this capacity to respond to signs of which we are not aware seems to guide our 'recognition of action as directed or purposive' It constitutes an instance of the very important statement that in many of the situations in our social world we understand without knowing how we understand.

> Whenever we conclude that an individual is in a certain mood, or acts deliberately or purposively or effortlessly, seems to expect something or threaten or comfort another, etc., we generally do not know, and would not be able to explain, how we know this. Yet we generally act successfully on the basis of such 'understanding' of the conduct of theirs. (S,48)[29]

4.2 Recognition

What kind of recognition is that? The basis of recognition is some sort of identification. But instead of taking this recognition as evidence of the existence of some essential common ground, Hayek warns that we have not so far arrived at any firm ground. That is, recognition is a problematic matter and shall be further explained. All that we may conclude is that our capacity of recognizing an action as rule guided relies upon our already possessing this rule. What is it that makes Hayek so

---

[29] On the role of physiognomic perception as a guide for practices and, more generally, of intuitive *Gestalt* perception of patterns, Hayek refers, in a footnote, to an earlier work by Michael Polanyi, namely, "Knowing and Being", Mind,70,1961. On that occasion, Hayek launches an attack against behaviorism and its physicalism. I quote: "we must also take into account in explaining the effects of men's actions that they are guided by such perceptions. (...) We shall find that perceptions of this sort, which the radical behaviorists wish to disregard because the corresponding stimuli cannot be defined in 'physical terms', are among the chief data on which our explanations of the relations between men must be built." (S,54)

prudent at this point of the argument? Why does he not take recognition as an evidence of 'sharing'? We should recall here that for Hayek the locus that each individual occupies in the social space is to a large extent unique; it is something resisting simplification of any sort, full comparability, or commensurability. So, if we recognize some action of others as meaningful it is because it matches our own interpretation of the situation that they face and, accordingly, seems to follow the applicable rule:

> This 'knowledge by acquaintance' presupposes therefore that some of the rules in terms of which we perceive and act are the same as those by which the conduct of those whose actions we interpret is guided. (S,59)

All we may say is that

> we must be made up of the same ingredients, however different the mixture may be in the particular instances. (S,59)

In this regard, Hayek advances in his The Counter Revolution of Science the idea that individuals in social interaction are "merely the *foci* in the network of relationships". This is an idea that seems to encompass a quasi-constitutive notion of expectations, for in a sense it is the way individuals are looked upon by others which affirm what they are doing/intending etc. So, in understanding what an individual is doing or intending to do or believes, we should not be concerned with the totality of his mind in all its complexity. In particular,

> if the social structure can remain the same although different individuals succeed each other at particular points, this is not because they succeed each other in particular relations, in particular attitudes they take towards other people and as the objects of particular views held by other people about them. The individuals are merely the **foci** in the network of relationships and it is the various attitudes of the individuals towards each other (or their similar or different attitudes towards physical objects) which form the recurrent, recognizable and familiar elements of the structure. If one policeman succeeds another at a particular post, this does not mean that the new man will in all respects be identical with his predecessor, but merely that he succeeds him in certain attitudes of his fellowmen which are relevant to his function as policeman. But this is sufficient to preserve a constant structural element which can be separated and studied in isolation. (CR,34)

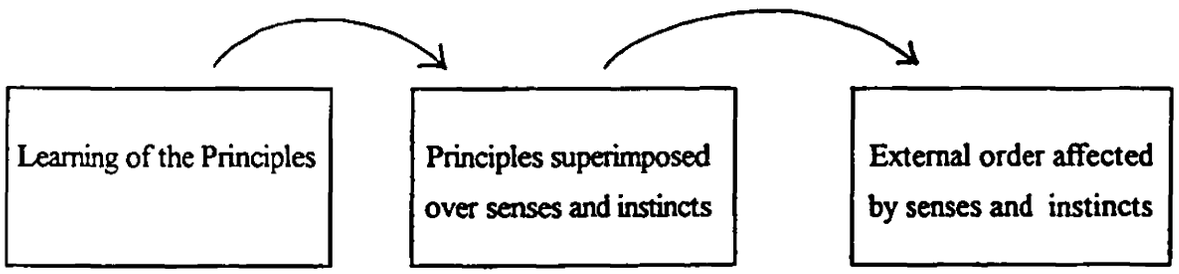
But again, to conclude that recognition amounted to 'function' would be an impoverishment of Hayek's broader intuition. For he acknowledges the complexity surrounding the arrangement of the superimposed and interrelated rules:

> The complexity of the arrangement in which these rules may be superimposed and interrelated is difficult briefly to indicate. We must assume that there exists not only on the perceptual side a hierarchy of superimposed classes of classes, etc., but that similarly also on the motor side not merely dispositions to act according to a rule but dispositions to change dispositions and so on will operate chains which may be of considerable length. (S, 57)

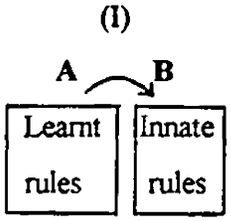
Additionally, this complexity is set in motion in a particular action when both a combination of the rules may be in order and a change in the usual way of dealing with it:

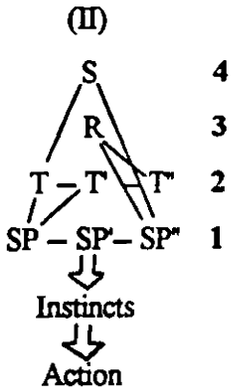
> It is this determination of particular actions by various combinations of abstract propensities which makes it possible for a causally determined structure of actions to produce ever new actions it has never produced before, and therefore to produce altogether new behavior such as we do not expect from what we usually describe as a mechanism. (NS, 49)

In The Fatal Conceit Hayek suggests that as far as a concept of order in the external world is concerned we should think of a scheme like the following:



Of course the process of learning may vary enormously and accordingly affect the whole chain. More generally, though, it is the complexity of the arrangement of principles itself which is at stake, which would have to include the multi-level rules we have been considering in the process of rule selection that an action sets forth. Consider the following diagrams:





As mind and society are not monocentric orders but only polycentric ones, problems of hierarchy and interrelation are more likely to arise and accordingly they will not always provide convergent solutions. We might agree that evolution has settled down the hierarchy **A – B**, as in (I), but certainly not any specific **4 - 3 - 2 - 1** hierarchy or combination in (II), as it would amount to denying the dynamic forces that evolution itself releases.

So, recognition carries the difficulty that we can never be certain of the interpretation we have given to an action of others, as we can never be certain that some particular fact actually corresponds to the place we have assigned to it within some pattern, as has already been remarked upon. That is, individuals are as complex structures as social interaction, their actions being "determined by the relation and mutual adjustment to each other of the elements of which (they) consist".[30]

This is the kind of uncertainty that 'knowledge by acquaintance' cannot get rid of, although it is true that the presumption that 'rules underlie people's actions' works for the improvement of the predictability of the environment. To be sure, action would falsify the local theories/expectations/interpretations held by agents because of the composition effects unless rules were followed. However, uncertainty still persists to the extent that we cannot be sure about the rules themselves.

---

[30] Cf. Studies.., p.73.

However, the very benefits we hope to derive from the presumption of a 'knowledge by acquaintance' stem from the fact that where there is some uncertainty, action is still possible for we may have the basis to attach 'a degree of confidence' to our interpretations.

The degree of confidence we attach to the interpretation we have made of someone's action is a function of the distance between us.[31] If we are quite distant from one another the reliability of any physiognomic knowledge is quite low. In this case, we do better to stick to more general and abstract rules to understand each other's actions, such as customs, habits and tradition.

Again, we may behave according to norms because this makes the consequences of our actions more predictable. Of course to make sense of this statement we have to consider that there are lower and higher level norms, that is, norms we abide by in a strategic sense (e.g., by deliberately following a habit which gives prospects of success) and others that are operating behind our backs, like *Sprachgefühl*. We might opt for predictability in the sense of actively following the lower level norms, but this is less clear with respect to the higher level norms. We may wonder whether higher level norms, less likely to be used in a strategic sense, will still encompass more contradictory guides to actions and accordingly lead to conflict.

We should note, then, that according to Hayek to follow a rule is also a way towards predictability, though not a prospect for a well informed utility maximizing choice:[32]

---

[31] Schutz makes a similar point, when he refers to the extent to which we resort to 'ideal types' when going about the social relations as a function of what he calls the 'degrees of anonymity' or indirectness in these relations. We quote from Wagner's comments: "indirect relationships fall into extended continuum patterns of growing anonymity. They range from the region of persons I have encountered and may encounter again to the artifacts which merely bear witness of the one-time existence of their entirely unknown makers." (ibidem,p.37) The degree of anonymity of the indirect relations determines whether I am 'We-oriented' or 'They-oriented'.

[32] The idea here is that even probabilistic calculations are ruled out, for they would be reliable just in an environment where repetition is likely to occur. Since Hayek deems our social world to be just circumstantially repeatable, not essentially, measures of expected utility are not applicable.

> Man does not so much choose between alternative actions according to their known consequences as prefer those the consequences of which are predictable over those the consequences of which are unknown. What he most fears, and what puts him in a state of terror when it has happened, is to lose his bearings and no longer know what to do. The world is fairly predictable only so long as one adheres to the established procedures, but it becomes frightening when one deviates from them. (S,80/81)

Still, this instrumental view of the rules does not fit well into Hayek's idea of normative knowledge as a constraint on actions. One conciliatory conjecture is to presume that the normative knowledge is made up of a mixture of internal (subjective) elements and externally given ones. We turn now to the 'internal' side of normative knowledge.

4.3 Addition

What is the basis of our understanding of what surrounds us, of our grasping the rule, under uncertainty? It seems that we understand, Hayek suggests, by **adding** things to the existing things, persons, objects around us. The image suggested here is that of the social actor as the superimposition of a quasi-material supplement gathered to him by the others and reciprocally by him to them. This material is 'opinions'.

Recognition is, according to Hayek, projection, 'reading into'. This is something Hayek makes quite clear when he analyses the meaning of what he calls 'social facts'. The objects of human activity (tools, food, medicines, weapons, words, sentences, communications, and acts of production)

> are defined not in terms of their 'real' properties but in terms of opinions people hold about them. In short, in the social sciences the things are what people think they are. (IEO,69)
>
> (...)
>
> People do behave in the same manner towards things, not because these things are identical in a physical sense, but because they have learnt to classify them as belonging to the same group, because they can put them to the same use or expect from them what to the people concerned is an equivalent effect. In fact, most of the objects of social and human action are not 'objective facts' in the special narrow

> sense in which this term is used by the Sciences and contrasted to 'opinions', and they cannot at all be defined in physical terms. So far as human actions are concerned, the things are what the acting people think they are. (CR, 27)

Still, if we want to understand someone's acts we shall then try to think what the observed person thinks she is doing, what her intention is. Yet, just as in the case of objects the understanding of which turns on what people's opinions about them are and if we want to know them we have to look at what people know about them, this is also so in the case of actions. Whenever we want to understand someone's actions as intentional or purposive we have no other choice but to **impute** this intention or purpose to the person on the basis of an analogy of our own mind (which we should think in terms of the same superimposition) and see if it works.

> On watching a few movements or hearing a few words of a man, we decide that he is sane and not a lunatic and thereby exclude the possibility of his behaving in an infinite number of 'odd' ways which none of us could ever enumerate and which just do not fit into what we know to be reasonable behavior - which means nothing else than that those actions cannot be interpreted by analogy of our mind. (IEO,64)

Recognition entails constraining, that is, limiting the range of alternative possibilities within which we can make sense of what others are doing, but also constructing. In 'consuming' normative knowledge we are also 'producing' more of it and adding to it, as our opinions and expectations testify.[33] This is done deductively, not inductively, Hayek contends:

> in discussing what we regard as other people's conscious actions, we invariably interpret their action on the analogy of our own mind (...) We thus always supplement what we actually see of another person's action by projecting into that person a system of classification of objects which we know, not from observing other people, but because it is in terms of these classes that we think ourselves. (IEO,63)

---

[33] That we are not passive recipients, determined by preexisting cultural ideas is also Schutz's view. I quote the comment by Wagner: "Schutz showed that even the socially most stereotyped cultural ideas only exist in the minds of individuals who absorb them, interpret them on the basis of their own life situation, and give them a personal tinge (...)". (ibidem,p.17)

We can never be sure about this imputed intention or knowledge, although this will not be cause for complaint as it may be sufficient for practical purposes. Moreover, this practice seems to be one of the sources of the realities (and conflicts) in our social world.

Another suggestive illation seems to be that this interpretative work, which is always taking place within social interaction, is analogous to the theoretical work of the social theorist, for both kinds of interpretation are fairly constitutive of their objects. A social theorist is only a more sophisticated observer than ordinary observing individuals engaging in daily interaction. These latter are also always acting on suppositions about the behavior of some others; by performing their decisions they generally display a high degree of confidence about the overall order that contains the farther frontiers of their acts. The social theorist, who works on second level suppositions, is bound to contain the variability of the world within some ideal types.

The similarity between action and theory is remarkable. Social theorists and actors are both building up their worlds based on suppositions on suppositions. In fact, Hayek contends that social sciences

> are not **about** the social wholes as wholes; they do not pretend to discover by empirical observation laws of behavior or change of these wholes. Their task is rather, if I may so call it, to **constitute** these wholes, to provide schemes of structural relationships which the historian can use when he has to attempt to fit together into a meaningful whole the elements which he actually finds. (IEO,72)

And we may extend the analogy to the actor in the social world as the actor 'constitutes' the social world within which he acts:

> The question is here not how far man's picture of the external world fits the facts, but how by his actions, determined by the views and concepts he possesses, man builds up another world of which the individual becomes a part. And by 'the views and concepts people hold' we do not mean merely their knowledge of external nature. We mean all they know and believe about themselves, about other people, and about the external world, in short everything which determines their actions, including science itself. (CR, 24)

All in all the issue of imputation seems to preserve a distance between the *explanans* (social interaction) and the *explananda* (rules), for the implication is that each actor works on **presumptions** concerning others' actions and whenever they prove not to work a change is set in motion.[34]

One kind of 'constitution', however, is ruled out by Hayek. In discussing the nature of the 'facts' in the Social Sciences, Hayek points out that our facts should be the 'ideas' held by people we, the theoreticians, are observing. At this point he distinguishes two levels of ideas that we should consider whenever we analyze social life, the one being opinions and beliefs ordinarily held by people about some object of their interest, which are constitutive of the whole called 'society', and the other being those speculative concepts that amount to popular theories or generalizations from the concrete knowledge people may have (that is, that 'normative' knowledge which has an inductive origin in explanatory knowledge).

> The beliefs and opinions which lead a number of people regularly to repeat certain acts, e.g., to produce, sell, or buy certain quantities of commodities, are entirely different from the ideas they may have formed about the whole of the 'society', or the economic system, to which they belong and which the aggregate of all their actions constitutes. (CR, 37)

These latter may be wrong and a major task for the social theorist it is to improve on them with the aid of the compositive method.[35] In fact, Hayek warns against the danger of mistaking these ideas for the 'facts' of our social sciences.

---

[34] This seems to fulfill one of Turner's criticism, that against the assumption of 'internal sameness' of the actors generally implied in the approach 'Practices-Tradition-Tacit Knowledge'. According to Turner, this approach does not keep the necessary distance between the *explanans* (order) and the *explananda* (practices, tradition and so on) when they explain order through practices, for they do not consistently display how **different** people came to acquire the **same** practices which are deemed to be conducive to order. See Turner (1994).

[35] Hayek proposes the compositive method in his CR

We may infer from this digression that speculative reasoning might improperly compete with established rules on the level of practical life. Given the ineradicable synoptical illusion entailed in the generalizations that people make (always from their ineradicably particular positions), Hayek finds a ground for rejecting (all too readily) any positive, constitutive function of these concepts whatsoever. These latter would release entropic forces within a spontaneous order, or so it seems. This is more difficult to make sense of. We shall turn to this point later in sections 6 and 7.

5. Order and disorder

So far we have come to the conclusion that every action in our social world is somehow *ex-ante* interaction, in the sense that it is based on our following rules and it is performed on the basis of a quite subjective interpretation of others' behavior as rule-guided as well. Yet we should not be misled into concluding that conformity to an orderliness is an inevitable outcome of the interplay between these interacting rule-following individuals. The positional differences, so to speak, among the individuals can be a source of both orderliness and disorder. We may provisionally refer to two sources of differences: the differences that might come about from different positions of the parts relative to the **whole**; and those which might arise from different positions of the parts relative to the other **parts**. So we have to spell out further conditions that might provide for orderliness. These will be the elements that contribute to the characterization of the autonomy of the whole with regard to its constituting parts.

Again, though this common background of rules was what made it possible for an orderliness to emerge from the interaction of the many individuals over time, it is important to note that orderliness is not a necessary achievement of the interaction itself, according to Hayek. Rule-guided interacting individuals may also produce an overall disorder. In other words, a system of individual rules may produce either order or disorder:

> Not every system of rules of individual conduct will produce an overall order of the actions of a group of individuals; and whether a given system of rules of individual conduct will produce an order of actions, and what kind of order, will depend on the circumstances in which the individuals act. The classical instance in which the

> very regularity of the behavior of the elements produces 'perfect disorder' is the second law of thermodynamics, the entropy principle. It is evident that in a group of living beings many possible rules of individual conduct would also produce only disorder or make the existence of the group as such impossible. A society of animals or men is always a number of individuals observing such common rules of conduct as, in the circumstance in which they live, will produce an order of actions. (S,67)

In other words, systems of rules may produce two different effects, order or entropy. Besides, an overall order of actions may be the result of different sets of rules of individual conduct. The same effect may result from different causes. What characterizes structures that permit an overall order to appear, then, is that they not only spring from a system of rules but also that these rules have somehow enabled under the circumstances a certain relationship between the parts and the whole which, in turn, conveyed a high degree of **autonomy** to the structures themselves:

> The orderliness of the system of actions will in general show itself in the fact that actions of the different individuals will be so co-ordinated, or **mutually adjusted** to each other, that the result of their actions will remove the initial stimulus or make inoperative the drive which has been the cause of activity. (S,69, my emphasis)

So, one important element is the emphasis that is being put on the co-ordinating effects of an action rather than on the aspect of individual initiative it contains. It is still true that each part produces the whole but it is no less true that it does so through the effects that its actions produce on the others and on the overall, its very existence turning on the results it causes in this external world made of all the others and of everyone.

The process that is launched can be described as follows: some circumstances may render a practice more conducive to an order of actions; this may be a motive for this practice to persist, as each acting individual will tend to conform to it on his way to achieving particular ends, plausibly by imitating others who successfully followed it.

Let us try, now, to spell out the elements that make for the autonomy of the outside world. Some auxiliary ideas we need to work out are the ideas of complexity, organized complexity and the statement that the whole is more than the sum of its parts.

Societies are complex structures, or structures with **essential complexity**, that is,

> structures whose characteristic properties can be exhibited only by models made up of relatively **large** numbers of variables. (NS,26, my emphasis)

Societies, moreover, are structures exhibiting a rather **organized** complexity:

> [this] means that the character of the structures showing it depends not only on the properties of the individual elements of which they are composed, and the relative frequency with which they occur, but also on the **manner** in which the individual elements are connected with each other. (NS,27)

So, on the one hand we have the question of a large number of variables, the individual parts and their relative frequency - which makes a mechanical treatment of this structure quite implausible. (And possibly also excludes a probabilistic treatment, for repeatability is not a necessary feature of the phenomena within this structure as we have already mentioned.) On the other hand we have the question of the way the parts adjust. This refers to the local-coordination among the parts.

Still, these complex structures, such as societies, individuals, language, and law, are orders that can be said to be **more** than the totality of the regularities observable within their parts ('and cannot be reduced to them'). They exceed the sum total of local-coordinations. The idea here is that not only are the irreducibly many elements within this whole related in a particular manner but also that there is a peculiar 'interaction of the parts with an **outside world both of the individual parts and the whole**'. This must be an **order-enhancing** interaction. That is, horizontal or local co-ordination must have a global effect:

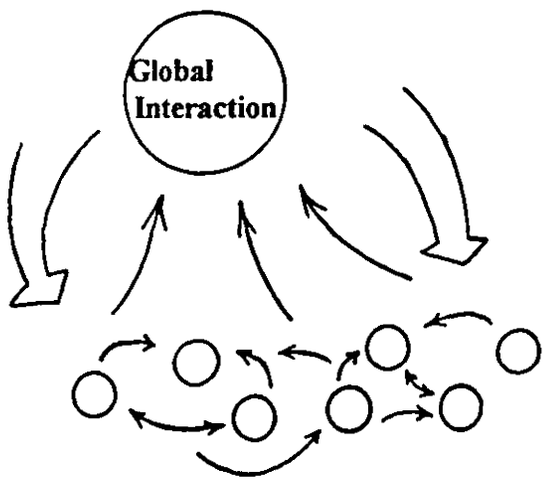
> If there exist recurrent and persistent structures of a certain type (i.e., showing a certain order), this is due to the elements responding to external influences which they are likely to encounter in a **manner which brings about the preservation or**

> **restoration of this order**; and on this, in turn, may be dependent the chances of the individuals to preserve themselves. (S,71, my emphasis)

There must have been a certain interplay among the parts and the outside world if there exists order, in particular that which proved somehow to benefit the parts and the whole simultaneously, and the part-in-relation-to-the-whole, through the effects the part itself produces on the whole. So, for it to be successful with respect to individual's adaptation, co-ordination in the local dimension must be capable of preserving or even launching global co-ordination

> We have to deal here with integration on at least two different levels, with on the one hand the more comprehensive order assisting the preservation of ordered structures on the lower level, and on the other, the kind of order which on the lower level determines the regularities of individual conduct assisting the prospect of the survival of the individual only through its effect on the overall order of the society. (S, 76)

Diagram of a complex system:[36]



The resulting order is then submitted to the test of the circumstances, to selection, that is. It should be noted in this regard that Hayek's notion of selection has a cultural connotation. He utterly rejects social Darwinism and claims that, with respect to human societies, we should be concerned with cultural rather than biological evolution and, accordingly, he approaches it with a Lamarckian perspective. This essentially means that for him, cultural evolution bears on the inheritance of

[36] This diagram is taken from Lewin (1993), p.13.

acquired characteristics - learnt rules guiding the 'inter-individual' relations - which are basically transmitted through external imitation.[37] To be sure, transmission is individual but selection operates on the group in that it is supposed to select an order of actions emergent from a system of rules which proves to adapt to external circumstances.

In this sense, selection cannot operate through falsificationism, for defeated rules can revive under favorable circumstances. Recall that different systems of rules may lead to order:

> systems of rules will develop as wholes, selection process of evolution will operate on the order as a whole: so, a new rule of individual conduct which in one position may prove detrimental, may in another prove to be beneficial (...) (S,71)

This fact may generate an endogenous movement, an internal testing to other rules:

> (...) and changes in one rule may make beneficial other changes, which before were harmful. (S,71)

However, the nature of this potential change is not really deeply examined by Hayek, as it is sometimes put on a par with spontaneous genetic mutation. But given that circumstances are ever-changing in a quite unexpected way, changes in the whole are highly likely to occur. Moreover, the whole's adaptive responsiveness turns on there already having been changes in some parts of the overall order of actions. How have these formerly undesirable changes appeared in the first place? How have they managed to subsist under unfavorable circumstances? How were they able to survive until the moment of their ultimate success? And what about the extension of the change itself? Could a given change manage to survive which we suppose to challenge important features of the overall order of actions? If so, isn't it conceivable that this new rule by itself forces its way on the order? To summarize, a convincing account of change is clearly missing.

---

[37] For a list of differences between cultural and biological evolution, see Hayek's The Fatal Conceit, "Between Instinct and Reason", p. 25 specially.

Now, there is something startling here. On the one hand, the surplus that complex and organized structures may generate is associated with the interactional ability in generating an order-enhancing interplay, at least in the weak sense of 'order-preserving'. On the other hand, though, this surplus is identified with another thing, perhaps the opposite of the previous one, namely, the imbalance a circumstance may generate within these structures when, in affecting some parts, it launches a critical test of the rules, and this may generate a disequilibrium in the system of rules that may induce further and unforeseen changes. That is, an order-disturbing move. Is this surplus itself a capacity to produce order or change? In conclusion, any overall coherence of the system of the rules, any compatibility with what we may call our tradition, is urgently calling for further vindication. Two clues apply: to get rid of consistency as the critical test, and adopt a dynamic image of order. In any case, a crucial question remains of how we are to discern whether 'disorder' (inconsistency) is preparing change or else setting deleterious forces in motion In other words, we need to differentiate 'good' inconsistency from entropy. It is time to examine the Hayekian concept of order.

6.Prospects for the third world

6.1 Tradition: cloudy horizon

Summing up, the elements that yield an autonomous arrangement are its essential complexity (composed by many variables), its organized complexity (it gives rise to local adjustments among its parts), and, finally, the fact that as a whole it exceeds the sum total of its constituting parts (local adjustments produce a relatively larger global adjustment).

Now from his epistemological premises alone Hayek cannot tell us whether an arrangement is 'order'. However, he does tell us so. He speaks rather vaguely (and naturalistically, as it were) about 'adaptation' or 'adaptive capacity' and 'group survival'. These are, nonetheless, very difficult things to ascertain, as we have seen from Hayek's approach to the system of the rules, unless one has some preconceived naturalistic idea of 'order' in mind, namely in terms of physical features such as quantities of population and wealth.

So, the implication is that Hayek seems to be taking the 'surplus' that a complex arrangement of individuals' actions produces not only as an epistemological one (that combined actions of the parts result in richness and unpredictability of the overall actions), but also as a 'moral' one (that the overall actions are a better suited arrangement), where this moral is expressed in a quite naturalistic fashion in terms of the whole's capacity for survival and the individual's ability to adapt to it.

I identify at this crucial point a tension between what I call Hayek's **evolutionary** approach and his more compromising **evolutionist** inclinations.

By his evolutionary approach I mean those interesting pieces of his epistemology that we have so far focused on, where the idea of evolution is worked out in terms of a way of generating knowledge, individual and interactional, in our social world. That is, Hayek's suggestion that we should think of our world as a changing structure, and change as having an important root in ongoing intersubjective relations; and, as a consequence of these findings, Hayek's conclusion that predictability and design are quite limited in our world.

However, Hayek surpasses these boundary conditions of a social order that are rooted in our knowledge predicament, which he can quite safely assert - that our world is composed of an array of intricate acts of production and consumption of knowledge - with a more substantive evolutionist view rather at odds with the predicament, where a sort of 'law of evolution' is implied, where social arrangements are succeeded by increasingly better ones, where the improvement is given in terms of certain 'efficiency' attributes.

To be sure, the concept of evolution is very unclear in this respect. As Lewontin (1968) reports, the concept of evolution entails, to varying degrees, some combination of the following elements: change, order, direction, progress and perfectibility. Change enwraps the idea of evolution in its simplest form:

> The idea of evolution, in its simplest form, is that the current state of a system is the result of a more or less continual change from the original state. The qualification that the change be continual, or at least frequent or regular, is an essential one and distinguishes the evolutionary from static world views. (Lewontin,1968:203)

As for order, this is more difficult to assert. How can we know whether an arrangement of the elements constitutes some 'order'? Should we accept any such arrangement as so? This is a matter that has divided those of evolutionary persuasion, as some have included direction, progress and perfectibility as features of an orderly state whereas others have been more reluctant in this respect.[38] In terms of our interest here, all we can say is that there is nothing in the Hayekian epistemology that authorizes him to establish a normative dividing line. However, Hayek cannot help displaying his normative preferences in terms of his views of direction, progress and perfectibility, the other seemingly more disputable elements of the concept of evolution which are, nonetheless, frequent in his later The Fatal Conceit.[39]

The evolutionary approach teaches us something about the epistemic foundations of our social world; it may warn us about some limits that this foundation imposes; but it cannot predict nor can it justify any substantive view of order on its own epistemological terms as an evolutionist view would require.[40]

In his more substantive idea of order, change has then a direction in time, in terms of achieving an ideal of progress and perfectibility, whereas his more formal account works out a sort of ontology of the social dimension, based on the premise of human ignorance alone.

---

[38] See Lewontin, op.cit.

[39] For a discussion of Hayek's defense of liberal principles in evolutionist terms, see Espada (1996), who claims that Hayek's view, for that matter, is both morally and logically untenable.

[40] We may trace some considerations to the effect that liberal societies have a comparative advantage over non-liberal ones to the extent that the former are structured as complex systems and, hence, allow a larger amount of knowledge to be freely disposed of through the operation of free markets, as Espada (1996) suggests. However this is not to demonstrate the absolute superiority of these arrangements. In Espada we may find a forceful criticism of Hayek's evolutionist view and the instrumental arguments issued therefrom in favor of liberal societies.

As an example of this more substantive view, a circular and rather static account of selection as an efficiency-searching method describes the relation between individuals, rules and overall order as follows:

> Although the existence and preservation of the order of actions of a group can be accounted for only from the rules of conduct which the individuals obey, these rules of individual conduct have developed because the individuals have been living in groups whose structures have gradually changed. In other words, the properties of the individuals which are significant for the **existence** and **preservation** of the individuals themselves, have been shaped by the selection of those from the individuals living in groups which at each stage of the evolution of the group tended to act according to such rules as made the group more **efficient**. (S,72, my emphasis)

However, as Hayek admits, a circumstance may cause a hitherto followed tradition to be less reliable than a required change in the rule. The meaning of order is less clear, then, for it is not clear **which** order of actions is being preserved or enhanced thereby.

We might think in terms of degrees of 'orderliness'. Still, what is the minimum degree of orderliness of a structure, which would allow us to characterize it as 'order'? How much orderliness is necessary to be able to characterize as successful, a complex structure? How much disorderliness is a society able to bear and still stay alive, persisting and reproducing?

Action displays a practical knowledge of 'what order is about', but, at the same time and by the same token theoretical knowledge is forbidden. We might **depict** order as a spiral movement which runs faster in the core than in the boundaries, with decreasing speed as long as we move from its heart to its peripheries. Dynamic images seem to be more suggestive.

A spiral-account sets evolution in time:

> though social theory constructs social orders from the rules of conduct assumed to be given at any one time, these rules of conduct have themselves developed as part of a larger whole, and at each stage of this development the then prevailing overall

> order determined what effect any one change in the rules of individual conduct had. (S, 73)

But it still sounds like a comparative-static analysis rather than dynamic in that we are interested in making sense of what happens **in between** one stage and the next, in understanding the nature of change and its impact, the crucial element of a complex order. Besides we are not so much interested in how a prevailing overall order determined the effect of a change in rules as how this latter affects the overall order of actions itself through its impact on other rules. Thus far, all we know is that every action inside the overall order may set in motion conservative as well as challenging forces with regard to the existing rules. The circumstances which will make a rule more likely to prevail than competing others, and the effect this will generate on other rules cannot be anticipated. What a miserable state!

Should we draw any conclusions regarding possible social theories? Beyond the poor predictive power of our theories that this picture of social world entails, we may hope that they may contribute to enhance the predictability of our world by influencing it on a modest basis, as Hayek suggests. But what are the means we may count upon in trying to achieve any modifications in the social world? Our social theory would scarcely be helpful. For a theory of social order will remain a theory of a number of ideal types as it cannot overcome the surprise of circumstances. In fact, it can say almost nothing in a deductive manner as to the results of the actions of the parts, of which particular facts the theoretician knows even less than the individual actors themselves. That is, though we are all deductivists we cannot make inferences from our premises, for they are to a large extent hidden.

Since ignorance is widespread in the world of suppositions and imputed knowledge that we live, in which, that is, nobody knows who knows best what we would like to know,

> the only way by which we can find out is through a social process in which everybody is allowed to try and see what he can do. (IEO,15)

Ignorance (at the individual level) generates knowledge (at the social level) and makes institutions and rules of conduct applicable, again as essential supplements to our self- and reciprocal-understanding. Still, individual ignorance is also a source of testing and potential challenge to these rules. But, what rules should we keep, then, in order to test any others?

What about the constructivist pretense that we may articulate our common premises and perfect our social order accordingly? Against it, there is the argument that what we more reliably have in common is inarticulable, known by acquaintance but not by understanding. Then there is the argument that our sense of justice is one instance of this unarticulated knowledge which may be represented in our laws, so even these cannot help relying upon some unarticulated premises. A more fundamental argument is provided by the idea that being the premises of our thinking and acting highly inarticulable we can only conjecture, interpret, and in doing so it is quite unlikely that we may overcome our position within the whole, with regard to that whole (what we know and do not know of what all the others know) and with regard to that part we are directly playing (our perception of what relevant others are doing). Rationality will appear more as an interpersonal and evolving capacity, than as a singular and atemporal attribute of an individual.

So, if constructivism is not convincing and tradition is not reliable, are there any outlets?

6.2 Third World

I shall insist on tradition,[41] the element Hayek identifies as enabling an overall order to come about and persist. We have seen that tradition is something that draws us towards conformity as well as encompassing a potential for change. It has been said that rules evolve, by a mechanism of transmission among the individuals and selection, by the whole, of those rules whose prospects of group success are the greatest. Yet, when we think of tradition we can think of a number of different references not thoroughly compatible with one another, nonetheless persisting.

---

[41] Tradition, here, stands for the broad system of largely unarticulated rules. I shall afterwards, in a later section, justify our choice of tradition with the help of the Popperian contention that a comprehensive traditionalism, so to speak, is logically tenable. I then refer to a particular tradition, that of critical rationalism. See section 7.

We may think in particular of some traditions as disappearing and reappearing at the mercy of circumstances, Hayekian circumstances which may render a rule beneficial after being considered detrimental. Or we may think, in a Popperian vein, of practices or rules not applicable to the present but by merely offering themselves as alternatives to prevailing rules, they end up by producing the very circumstances under which they shall be considered applicable. We may also think of a number of different interpretations of our traditions (like freedom, and equality, and tolerance), as authorizing many different and even contradictory rules and constraints.

Without pushing this point too far now, I would like to work out Popper's (1949,1987) suggestion that one of our main traditions is the very idea of reason as a critical capacity. Is Hayek's selective mechanism compatible with critical rationalism? Let us suppose them to be so, on the presumption that Lamarckian evolution - 'what is acquired may be transmitted' - is congenial with the invention of favorable circumstances for a change - '... and may be made to work under a newly established environment' So, I shall presume, critical rationalism is compatible with a phenotypical account of social interaction.

This would entail considering selection not as a thoroughly blind mechanism whereby our environment selects us, our practices and institutions, but as also active,[42] in that we would also have the capacity to select the environment, to influence our circumstances. Are the Hayekian actors capable of creating 'ecological niches', in the words of Popper, guided by their imagination, and taking advantage of the very mechanism whereby a spontaneous order operates?

Hayek explicitly admits the possibility that we may make the most of the operation of a spontaneous order. Yet would he be willing to support the Popperian third world, made up of

---

[42] Espada, op.cit., makes a similar point, in his defense of an 'indirect' and 'negative' constructivism, whose roots he finds in Popper's "In Search of a Better World".

inventors (though without an architect)? Let me try to spell out, in a provisional and quite desultory way, some elements of a possible conciliation.

One first element would be the Hayekian idea that social actors are (unsuccessful) constructivists. They act from their theories and opinions, beside the higher level supra-conscious rules which nonetheless undergo a non despicable interpretative work on the part of the individuals. The question arises, quite naturally it seems to me, as to the scope of consciousness in such individual undertakings.

The suggestion is that we should think of consciousness not only as one more complicating element behind our actions along with the unarticulated ones but also as a special element with a special force as it may take us through a long problem-solving path.

Secondly, we would have to approach 'circumstances' in a less natural or external way than is sometimes suggested in certain vetoes that Hayek utters. That is, at least and with Hayekian authorization, if the circumstances that constrain an action are to influence the individual's choice, they have to be **perceived** by the actor through the filters he makes use of, primarily through his perception of the past and of the future, the latter referring to the unavoidable anticipations he makes. So, the circumstances are to a large extent non-natural, but 'made', and, I conjecture in addition, not only passively made, that is as a side product of our trying to grasp what is going on led by our inarticulate guidelines, but also actively.

In this respect, Hayek discusses at length the significance of the distinction 'natural versus artificial'. He regards this latter as meaningless, for he considers that our world is man **made** (and in this sense non-natural), but also **undesigned** (and in this sense non-artificial). I am here working out the seemingly more suggestive distinction between a passive and an active manufacture. My suggestion is that our order is made up of weak-constructivists, who passively and actively contribute to its make up. But how? To what extent can man actively create his circumstances?

Hayek concedes that we may derive some help from our social theories, in the sense that they may influence the order by taking advantage of its spontaneous motion. Actually, Hayek contends that part of our theoretical efforts should be devoted to discerning what the rules of conduct are which might deliberately be improved upon. I feel, however, that this suggestion is not fully coherent with the limited place Hayek made abundantly clear we should assign to social theory. This being due to our limited capacity of understanding and more generally to the limited reach of reason.[43]

I shall conjecture, with Popper, that circumstances may operate as a selective mechanism (as in the idea that the environment selects us) but also through **criticism** (as in the idea that we can also select the environment). And the meaning of this criticism should exceed the mere test of consistency of existing rules and encompass an imaginative effort:

> we...owe to the third world...our rationality - that is, our subjective mind, the practice of critical and self-critical ways of thinking, and the corresponding dispositions. More important than all this...is the relation between ourselves and our work, and what can be gained for us from this relation.
>
> The incredible thing about life...is just this interaction between our actions and their results by which we constantly transcend ourselves, our talents, our gifts.
>
> This is how we transcend our local and temporal environment by trying to think of circumstances **beyond** our experience: by criticizing the universality, or the structural necessity, of what may appear...as the 'given' or as a 'habit'; by trying to find, construct, invent new situations - that is **test** situations, **critical** situations; and by trying to locate, detect, and challenge our prejudices and habitual assumptions.
>
> This is how we lift ourselves by our own bootstraps out of the morass of our ignorance; how we throw a rope into the air and then swarm up it - if it gets any purchase, however precarious, on any little twig. (Popper,1974:147/148)

Hayek stresses, it seems, the passive productivity of our social world whereas a Popperian argument about objective knowledge encourages us to portray this productivity as **also** active; a

---

[43] See Kukathas'(1989, p.82) point: "The contradiction into which Hayek seems to be locked is one in which he is forced to claim that there can be no social theory that can guide our evaluation of the entire structure of our institutions, while conceding that a considerable social theory, explaining the purpose served by our institutions, is necessary for us to evaluate and alter our rules of conduct."

more positive role is assigned to consciousness in this latter. The intra-personal aspect of 'over-production' has been underemphasized, in Hayek's more substantive view, in favor of an over emphasis on the interpersonal issue. This asymmetry has not caught Hayek's attention, for at the intra-personal level consciousness is conceivably to play a quite important role, having no analogue at the level of inter-personal relations. Hayek should have explored the consequences of it on the overall order, instead of dismissing it on the basis of its irrelevance at the inter-personal level (or the dangers plausibly resulting from the supposition of a collective consciousness). Though we may agree that individual consciousness is not able straightforwardly to construct or evaluate the overall order, it may, however, complicate the landscape as it may also contribute to change, thus actively changing the very parameters of order itself. The extent of this contribution still requires explanation.

Thus far, then, I have found no justification to Hayek's recommendation that we ought to be actively ignorant. The Hayekian overemphasis on a rather passive consciousness reveals Hayek's rejection of what he calls the 'abuse of reason', typical of constructivist thinking. For him, the rationality postulate, in taking for granted precisely what we should be more interested in understanding, begs the question: what can we know? But he should concede that to draw the full consequences of the epistemological conditions of individual action that he sets out, would amount to enlarging a bit the scope of consciousness as well as enlarging its meaning, that is, as a capacity of both foresight and hindsight, and foresight not as 'guesses' but as invention. Invention should be allowed, alas, precisely for the very fact that guesses are not possible under the ignorance condition and anyway we need to accommodate consciousness somewhere! At least he might concede that his world is populated by constructivists, who may not know their premises but might on purpose **add** some to their provision, for the capacity of addition is already recognized by Hayek himself when he discloses the reasons for which action is to subsist under radical ignorance.

If this is so, Hayek who fears feelings of ours like 'feeling the ground withdrawn from under our feet' can be made to support an argument which actually hopes that we may be able also on purpose, though not fully anticipating the outcome, to raise ourselves up by our own bootstraps.

7. Praxis and poiesis

> *"We are afraid to put men to live and trade each on his own private stock of reason, because we suspect that this stock in each man is small, and that the individuals would do better to avail themselves of the general bank and capital of nations and of ages."*[44]

In order to explore in a less desultory way some of the suggestions I have so far been making other voices shall be introduced into the debate. The guests are Michael Oakeshott, Karl Popper and Michael Polanyi. They are introduced at this point to pave the way to a more acceptable idea of tradition, more in line with Hayek's evolutionary (and not evolutionist) aspects. Let me begin with Oakeshott, who, to my purposes here, will represent a strong case against prescription and conscious intervention in the world of practical life.

According to the British historian Michael Oakeshott, politics and scientific activity, along with cooking, stumble upon the ineffable: these activities are emblematic of a sort of impossibility of rationality in that they make use of a practical (concrete and skillful) knowledge that cannot be reduced to a technical (abstract and ruled) knowledge, of which they also make use. However, the rationalist in politics, he complains in "Rationalism in Politics", will always try to assimilate politics to engineering:

> In this activity the character which the Rationalist claims for himself is the character of an engineer, whose mind (it is supposed) is controlled throughout by the appropriate technique and whose first step is to dismiss from his attention everything not directly related to his specific intentions. This assimilation of politics to engineering is, indeed, what may be called the myth of rationalist politics. (Oakeshott,1974:5)

The Rationalist's misleading doctrine about human knowledge is, in a few words, that 'more and more certain knowledge about men and society can be reached through the unencumbered

---

[44] Edmund Burke, Reflections on the Revolution in France, p.99.

intellect'.[45] The fact of the matter, says Oakeshott, is that every practical activity requires two sorts of knowledge, namely, the technical one which is constituted by rules susceptible of precise formulation, and the practical which only exist in use, is not reflective and cannot be formulated in rules. This latter being conveyed not through any method of formulated doctrine but through tradition:[46]

> Even in the **what**, and above all in diagnosis, there lies already this dualism of technique and practice: there is no knowledge which is not 'know how'. (ibidem:9)

Closer inspection reveals that Oakeshott also implies a hierarchy between these kinds of knowledge in that some practical knowledge of anything is always supposed to be the nourishing element for any technical knowledge to flourish. Indeed, to learn is to reform something already in existence:

> As with every other sort of knowledge, learning a technique does not consist in getting rid of pure ignorance, but in reforming knowledge which is already there. Nothing, not even the most nearly self-contained technique (the rules of a game), can in fact be imparted to an empty mind; and what is imparted is nourished by what is already there. A man who knows the rules of one game will, on this account, rapidly learn the rules of another game; and a man altogether unfamiliar with 'rules' of any kind ...would be a most unpromising pupil. (ibidem:12)

Oakeshott evokes, for that matter, affiliation to Pascal's theory of knowledge:

> the only knowledge that is certain is certain on account of its partiality; the paradox that probable knowledge has more of the whole truth than certain knowledge.(...) The human mind, he asserts, is not wholly dependent for its successful working upon a conscious and formulated technique; and even where a technique is involved, the mind observes the technique 'tacitement, naturellement et sans art'. (ibidem:20)

What is more, the attempt to formulate the rules may even have deleterious consequences:

---

[45] Cf. Oakeshott, op.cit.

[46] Cf. Oakeshott, op.cit.

> The precise formulation of rules of inquiry endangers the success of the inquiry by exaggerating the importance of method. (ibidem:20)

The origin of the Rationalists' confusion is his assimilation of the mind to the faculty of 'Reason', as in his 'supposition that a man's mind can be separated from its contents and its activities (...) the mind as a neutral instrument, as a piece of apparatus.'[47] Oakeshott, in "Rational Conduct", displays a different conception of mind:

> Now, this mind I believe to be a fiction; it is nothing more than an hypostatized activity. Mind as we know it is the offspring of knowledge and activity; it is composed entirely of thoughts. You do not first have a mind, which acquires a filling of ideas and then makes distinctions between true and false, right and wrong, reasonable and unreasonable, and then, as a third step, causes activity. Properly speaking the mind has no existence apart from, or in advance of, these and other directions. These and other distinctions are not acquisitions; they are constitutive of the mind. (Oakeshott,1974:90)

As for the precedence of the activity in relation to the 'rational' purposes of its practitioners, the following passages are illuminating:

> It is an error to call an activity 'rational' on account of its end having been specifically determined in advance and in respect of its achieving that end to the exclusion of all others, because there is in fact no way of determining an end for activity in advance of the activity itself; (...) A cook is not a man who first has a vision of a pie and then tries to make it; he is a man skilled in cookery, and both his projects and his achievements **spring from** that skill. (ibidem:91)

In a even stronger statement, the activity is said to define questions as well as approaches towards solutions:

> It is the activity itself which **defines** the questions as well as the manner in which they are answered (...) (ibidem:97/98, my emphasis)

---

[47] Cf. Oakeshott, 1974, "Rational Conduct", p.86.

> Both the problems and the course of investigation leading up to their solution are already **hidden** in the activity (...) (ibidem:99/100, my emphasis)

A final point shall present Oakeshott's view on what 'rational conduct' (as opposed to a 'rational faculty') should be, namely, 'faithfulness' to tradition:

> the only significant way of using the word 'rational' in relation to conduct is when we mean to indicate a quality or characteristic (and perhaps a desirable quality or characteristic) of the activity itself, then it would appear that the quality concerned is not mere 'intelligence', but **faithfulness to the knowledge we have of how to conduct the specific activity we are engaged in**. 'Rational' conduct is acting in such a way that the coherence of the idiom of activity to which the conduct belongs is preserved and possibly enhanced. (ibidem:102)

The agricultural metaphor of the practical knowledge being a nourishing element for the technical knowledge to flourish is taken here as an indication of a fairly concrete limit to conscious ambitions to transcend anything, and for the argument we are going to make in this section this metaphor shall constitute our model of 'praxis'. By this it is meant the idea that the activity, any activity in which we are involved in our practical life, is always bigger than its product, the result of the activity itself being always but an abridgment of its unknown possibilities. An instance of this is provided by Oakeshott's statement that "**what** we do, and moreover what we want to do, is the creature of **how** we are accustomed to conduct our affairs."[48]

From this statement at least two things follow: the first is that no analysis of the product is able to reveal the complexity of the activity and thereby allow any person to master it, in the same way as the rational analysis of a delicious meal in the form of its recipe is unable to convey the necessary knowledge for it to be made by an inexpert cook ; the second is that any change that might possibly happen is deemed to be already inscribed in the activity itself, the individual or group being nothing more than its midwife and only insofar as they have been brought up within it.

---

[48] Cf. 1974, "Political Education", p.120.

The earthly metaphor is suggestively replaced, in "Political Education", by an even less enabling liquid one when it comes to politics:

> In political activity, then, men sail a boundless and bottomless sea; there is neither harbor for shelter nor floor for anchorage, neither starting-place nor appointed destination. The enterprise is to keep afloat on an even keel; the sea is both friend and enemy; and the seamanship consists in using the resources of a traditional manner of behavior in order to make a friend of every hostile occasion. (Oakeshott,1974:127)

We should then rely on the politician's skills and ability to handle the tools provided by tradition on his way to '**attending to** the general arrangements of a set of people' and not '**making**' these arrangements. This art cannot be spelled out by any political doctrine. Hence, political philosophy, after Oakeshott, must be just a history

> of the incoherences philosophers have detected in common ways of thinking and the manner of solution they have proposed, rather than a history of doctrines and systems. (ibidem:132)

And, frowning on Hayek's position concerning social theory, Oakeshott concludes that:

> neither 'principle' (on account of what it turns out to be: a mere index of concrete behavior) nor any general theory about the character and direction of social change seem to supply an adequate reference for explanation or for practical conduct. (ibidem:136)

In this line we should recall the famous passage, in "Rationalism in Politics", in which Oakeshott disagrees with Hayek, accusing Hayek's rationalism:

> A plan to resist all planning may be better than its opposite, but it belongs to the same style of politics. (Oakeshott,1974:21)

It seems that, in this statement, Oakeshott realizes Hayek's paradox of producing a political philosophy of the unpredicted consequences and unarticulated premises, a paradox I want to explore in the opposite direction than that proposed by Oakeshott. To this end, I want to introduce some of Popper's reflections. For though belonging in the same tradition of criticism of rationalism

as Hayek and Oakeshott, Popper is still a rationalist, or so he claims. What he is certainly opposed to is what he calls the rationalist's 'immodesty'.

In fact, in at least two lectures (1945;1949) in which he discussed the issue of rationalism, Popper took up the anti-rationalist challenge, as remarkably raised by Edmund Burke in his famous work on the French revolution, and, more recently, by the 'Cambridge historian Michael Oakeshott'.

Though recognizing the force of the challenge, in his 1945 lecture, Popper makes a distinction between Burke and Oakeshott's critiques: while Burke is taken as a thinker who developed an "analysis of the importance of that irrational power which we call 'tradition'", Oakeshott deserves a less enthusiastic welcome:

> Quite a number of outstanding thinkers have developed the problem of tradition into a big stick with which to beat rationalism. I may instance Michael Oakeshott, a Cambridge historian, a really original thinker, who recently in the Cambridge Journal launched an attack on rationalism. I largely disagree with his strictures; but I have to admit that the attack is a powerful one. (Popper,1949:121)

Though agreeing with the importance of tradition, following the Burkeian route, Popper sides with rationalism, thus rejecting the seemingly radical anti-rationalism of Oakeshott. I am a rationalist of sorts, Popper declares in his 1949 paper. In his 1945 article, he states even more sharply that, as far as the struggle rationalism versus irrationalism is concerned, 'I am entirely on the side of rationalism'. And he adds further:

> This is so much the case that even where I feel that rationalism has gone too far I still sympathize with it, holding as I do that an excess in this direction (as long as we exclude the intellectual immodesty of Plato's pseudorationalism) is harmless indeed as compared with an excess in the other. In my opinion, the only way in which excessive rationalism is likely to prove harmful is that it tends to undermine its own position and thus to further an irrationalist reaction. It is only this danger which induces me to examine the claims of an excessive rationalism more closely and to advocate a modest and self-critical rationalism which recognizes certain limitations. (Popper,1945:p.34)

But how to reconcile tradition and rationalism? First, Popper suggests, by accepting that we cannot free ourselves from traditions; second, by assuming a critical attitude towards tradition in the sense that we should not accept it as unquestionable taboos; third, by inscribing any critical attitude of ours in a tradition, the 'second-order' tradition of critical rationalism, of critically inspecting the ideas offered to us and offering our own ideas to critical scrutiny.

The only kinds of rationalism that do not pass this test are uncritical and comprehensive rationalism, namely, the idea that 'I am not prepared to accept anything that cannot be defended by means of argument or experience' (a logically untenable statement which itself is not grounded either in argument or experience), and the disregard of the fact that 'no rational argument will have an effect on a man who does not **want** to adopt a rational attitude' (a rational attitude depends on a pre-rational decision).[49]

Hence, although comprehensive rationalism is logically untenable whereas comprehensive **irrationalism** is, in addressing the pertinent points raised by Burke and later by Oakeshott, Popper launches a defense of rationalism that seemingly renders it logically tenable. He restates rationalism, as it were, as comprehensive **traditionalism**, and the tradition referred to is that of critical rationalism, the one invented (sic) by the Greek philosophers.

So, Popper's very peculiar form of rationalism, what he calls critical rationalism, is advanced on the grounds that criticism and not any substantive content is the crucial feature of reason, one which is deeply inscribed in our tradition. Rationalism is useful as a pair of spectacles for us to read our tradition but is itself just a tradition and therefore cannot be rationally understood. Hence, Popper's world contains both this messy thing called tradition, along with its articulable and non-articulable elements, and the critical glance which came out of it, which is, however, an attitude we may decide (on pre-rational grounds) to take.

---

[49] Cf. Popper, 1945, pp 34 and 35, respectively.

Popper's views on tradition, thus, seemingly reconcile usual antinomies such as 'tradition versus change', and 'tradition versus rationality', when approaching both scientific activity and society. As for the antinomy, tradition versus rationality:

> It is not possible for you to act rationally in the world if you have no idea how it will respond to your actions. Every rational action assumes a certain system of reference which responds in a predictable or partly predictable way. Just as the invention of myths or theories in the field of natural science has a function - that of helping us to bring order into the events of nature - so has the creation of traditions in the field of society. (Popper,1949:131)

As for tradition and change:

> The analogy between the role of myths or theories in science and the role of traditions in society goes further. We must remember that the great significance of myths in scientific method was that they could become the objects of criticism, and that they could be changed. Similarly traditions have the important double function of not only creating a certain order or something like a social structure, but also giving us something upon which we can operate; something that we can criticize and change. This point is decisive for us, as rationalists and social reformers. (Popper,1949:131)

We may think of the interplay between tradition and reason as a reciprocal limitation, where our critical capacities impede us from following traditions blindly. That 'we cannot start afresh' or 'clean the canvas' restrains our constructivist 'immodesty'. We may think of it also in terms of openness,[50] as in Popper's suggestion that we can 'invent' traditions. This latter possibility will constitute our model of 'poiesis'. By this is meant a model of practical life in which the result of a practical activity, its product, is bigger than the activity itself: it displays a surprising outcome that does not fit well into the traditional modes of behaving or understanding within the activity and thereby supports a genuine drive for change. The best example here is the already mentioned Popperian third world[51]

---

[50] Espada (1996) uses the same term 'openness' to state his point in favor of a more active role for social actors in the political society, in a different interpretation from the one we take up here, where 'openness' gives room to piecemeal intervention in a political society.

[51] I refer to Popper (1974).

of objective knowledge, and the extremist version of it may be seen in the following quotation selected from Vernon (1976):

> if we try, we can break out of our framework at any time. Admittedly, we shall find ourselves again in a framework, but it will be a better and roomier one, and we can at any moment break out of it again. (cited in Vernon,1976:273, quoted from Popper)

This quotation suggests that we might so enlarge the sense of reason as to encompass not only a capacity of effecting consistency or coherence tests and checking contradictions, but also an imaginative capacity that would let us transcend, at least in an intellectual sphere (Popper talks about scientific research), the world as given.

Vernon warns us to take care with the fact that when it comes to politics the sense of openness implied in Popper's (1974) poiesis is quite narrowed down when compared to his 'revolutionary' attitude towards our modes of thinking within the scientific community. Though acknowledging that openness is much more limited in Popper's thinking as far as political society is concerned, we shall not give up an examination of the consequences of some of Popper's premises concerning our conjectures, beliefs and expectations as long as these are also inhabitants of a political society.

In short, under the model of praxis the stress is being put on continuity whereas under the model of poiesis, the emphasis is clearly on the side of change. Somewhere in between the self-proclaimed 'conservative' Oakeshott and the 'revolutionary' Popper we should have to make room for an alternative perspective. To this end, we introduce Michael Polanyi's contribution. As we will see, Polanyi provides a firmer basis to Popper's poiesis.

I have already referred to Polanyi's The Logic of Liberty (1951), where the distinction between monocentric and polycentric orders is proposed and which exercised a most effective influence on Hayek's thinking. In the late introduction (1963) of his short book Science, Faith and Society (SFS,1946), Polanyi launched an attack against the idea of the planning of science that was then officially sponsored in the Soviet Union and gave rise to the subsequent enforcement of Marxist

philosophy, and began to spread to England under the form of a 'mental disturbance', where a number of British scientists began to defend ideas such as 'science for the citizen' and the like. His The Logic of Liberty dealt to a great extent with this issue, and SFS is written in the spirit of addressing this erroneous idea. He wanted to take up the following questions, challenged as he felt by Marxism:

> What philosophy of science had we in the West to pit against this? How was its general acceptance among us to be accounted for? Was this acceptance justified? On what grounds? (Polanyi,1946,1966:9)

In search for the answers, Polanyi declares:

> Like the Marxist theory, my account of the nature and justification of science includes the whole life of thought in society. In my later writings it is extended to a cosmic picture. But the ultimate justification of my scientific convictions lies always in myself. At some point I can only answer, 'For I believe so'. (ibidem:9)

Knowledge is dealt with as only 'perceived knowledge', much in the tradition of phenomenology; Polanyi refers to Maurice Merleau-Ponty's La Phenomenologie de la Perception (1945) as an analysis of 'perceived knowledge on the lines of Husserl'[52] expressing views akin to his own. His emphasis is on two elements, namely, the 'tacit coefficient of explicit knowledge' and the 'tacit process by which scientific knowledge is discovered'.[53] The first overrules the possibility of 'scientific control' over scientific knowledge, in the sense that the 'premisses of science on which all scientific teaching and research rest are the beliefs held by scientists on the general nature of things.'[54]The second, similarly overrules a control over the process of scientific discovery:

> there are rules which give valuable guidance to scientific discovery, but they are merely **rules of art**. The application of rules must always rely ultimately on acts not determined by rule. (...) The rules of scientific enquire leave their own application wide open, to be decided by the scientist's judgment. (Polanyi, 1946,1963:14/15)

---

[52] Cf. Polanyi, 1946,1963, p.12.

[53] Cf. Polanyi, op. cit., p.13.

[54] Cf. Polanyi, op.cit., p.11.

We shall pursue now, to a certain extent at least, Polanyi's ideas in his Personal Knowledge, where an original argument stresses the most important role personal deliberation is to play in a traditionalist perspective.

The key words of Polanyi's outlook in this work, for my particular undertaking here, are the notions of 'personal knowledge' and 'passive action'. It is certainly a huge task for anyone to sum up Polanyi's contribution to the understanding of human action, and of scientific practice in particular, as displayed in his Personal Knowledge. Fortunately, this is not my aim here. What shall interest me most are some of his insights and the possible effects they might generate on our proposed attempt to fill missing parts of Hayek's understanding of the social order emergent from interacting individuals. One crucial element is the notion of personal knowledge. Let us turn to it now.

Polanyi argues for the existence of a 'personal knowledge' on rather philosophical grounds, for it is something, he believes, that is able to bridge the gap between subjectivity and objectivity:

> We shall find Personal Knowledge manifested in the appreciation of probability and of order in the exact sciences, and see it at work even more extensively in the way the descriptive sciences rely on skills and connoisseurship. At all these points the act of knowing includes an appraisal; and this personal coefficient, which shapes all factual knowledge, bridges in doing so the disjunction between subjectivity and objectivity. It implies the claim that man can transcend his own subjectivity by striving passionately to fulfil his personal obligations to universal standards. (Polanyi,1958:17)

The synthesis personal knowledge (PK) represents may be felt in its capacity to bring together the particular and the universal elements that are present in our attempts at acquiring a knowledge of reality: we personally **strive** for **universal** standards, or for getting more and more intimations of reality. This sort of passion for the universal is the effective capacity we have to come closer to reality. And the evidence of any success we might have in this striving is given by the implications of our efforts going beyond the original experience our knowledge used to control. So, an index of

reality is spontaneity which can only be felt through a contrivance, a conscious act. In short, if we do not try to know more, we will not know anything, or just know less.

The roots of personal knowledge are two, according to Polanyi: focal awareness and subsidiary awareness. These are two different and mutually exclusive kinds of knowledge which are present in every practical activity of human life. Focal awareness is knowledge or action that can be reduced to rules, or cause-consequence relations; subsidiary awareness is the 'tools and probes' we assimilate when trying to attain focal awareness.

More generally, subsidiary awareness is the set of our pre-suppositions, the interpretative framework that enables us to be focally aware of something. So, a performance cannot be accounted for just on the basis of a logical analysis for it contains more than that, and in this sense full self-consciousness destroys the sense of context which alone can evoke the sequence of words, gestures, and so on. This refers to the unspecifiable character of some arts like scientific research and politics.

> The kind of clumsiness which is due to the fact that focal attention is directed to the subsidiary elements of an action is commonly known as self-consciousness.(...) This destroys one's sense of the context which alone can smoothly evoke the proper sequence of words, notes, or gestures.
>
> (...)
>
> We may describe such a performance as **logically unspecifiable**, for we can show that in a sense the specification of the particulars would logically contradict what is implied in the performance or context in question. (ibidem, 56)
>
> (...)
>
> All particulars become meaningless if we lose sight of the pattern which they jointly constitute. (ibidem, 57)

This is not new, it is just another way of saying that the order exceeds the sum of its constituent parts, and accordingly the understanding of the order is not tantamount to the analysis of its

elements. The interesting suggestion by Polanyi refers to the way we acquire this PK and the dynamics that takes place between its two different roots.

So, how is PK achieved? For Polanyi, it is achieved by a repeated mental effort at the instrumentalization of a thing to some purpose:

> Our subsidiary awareness of tools and probes can be regarded... as the act of making them form a part of our own body.... We may test the tool for its effectiveness or the probe for its suitability, e.g., in discovering the hidden details of a cavity, but the tool and the probe can never lie in the field of their operations; they remain necessarily on our side of it, forming part of ourselves, the operating persons. We pour ourselves into them and assimilate them as parts of our own existence. We accept them existentially by dwelling in them. (ibidem, 59)

So, in trying to learn/understand/solve something, which is a conscious and purposeful act we end up with two different payoffs, so to speak, one being an increase in our subsidiary awareness, the unconscious elements we assimilate in the process, and the other being a newly acquired consciousness which is tantamount to the thing we learnt or came to understand or solved and may even go beyond. So, this is a positive relation between conscious and unconscious acts whereby a conscious effort sets in motion a process of unconscious trial and error, not fully specifiable, by which we grope our way towards success. This is what Polanyi calls a **passive action**.

Still, the negative relation between articulation and non-articulation to the effect that not every thing which is known may be consciously known also gains, with Polanyi, a more fertile meaning than the ones suggested by both Hayek and Oakeshott. He seems to imply that this relation should be seen not only as the inarticulable parts of our knowledge constituting the condition of possibility of the articulable ones but also and more interestingly as an index of a non-abridgeable and indeterminate distance between both elements. This latter point may be made clearer through an illustration, as follows.

Polanyi takes the relation between speech and thought to give rise to three distinctive domains: the ineffable domain (fully inarticulable), the domain of co-extensiveness (as in the case of a text and its

meaning), and the domain of sophistication (where articulation and non-articulation are thoroughly unconnected). This latter domain requires a decision on the part of the recipient of the message, either in the form of a speech or a text: he has to decide whether the lack of meaning is due to a fumbling that might, henceforth, be corrected, or a pioneering, that is a proposed change in some of the elements that constitute meaning.

So, the relation articulation-non-articulation involves an indeterminacy, and, particularly in the case of language, that the symbolic operations cannot be fully understood is an insurmountable feature of language itself. According to Polanyi, this indeterminacy is double: we cannot say all that we know or think; and we cannot know or think all that is implied by what we say. We have then backwards and forwards indeterminacy. What does this mean? For Polanyi, indeterminacy may be taken as an index of an anticipatory power of language, as a signal of its contact with reality (Polanyi's words).

> I have already affirmed that these indeterminate anticipatory powers of an apposite vocabulary are due to its contacts with reality. We may extend the conception of reality implied here to account for the capacity of formal speculation to raise new problems and lead on to new discoveries. A new mathematical conception may be said to have reality if its assumption leads to a wide range of new interesting ideas. (PK, 116)

In short, the message Polanyi seems to be conveying is not the same as that uttered by Oakeshott, for it is not 'don't try to overcome the boundaries of the unarticulated, for you won't succeed (Oakeshott) or you will destroy its motto (Hayek)', but something else. What for him is more worthwhile is precisely the efforts we make to invade the unknown, to have an anticipation of the unexpected, to get an intimation of reality. And this is done through a passive action, a **deliberate** act that prepares the conditions for the **emergence** of a novelty. The example he gives of the operation of this passive action is scientific research where the whole activity of problem-solving may be seen as a contrived act which enables the emergence of a 'happy thought': heuristic efforts evoke their own consummation.[55]

---

[55] Cf. Polanyi, 1958.

So, Polanyi acknowledges a pretty important place for design within the practical world through his notion of personal knowledge which is in fact pointing to an non-eliminable coefficient of individual initiative in the mechanics of our world. This yields a more detailed basis for the Popperian poiesis.

Turning to our provisional typology of praxis and poiesis I will now try to place Hayek within it. First of all, we should stress Hayek's concern with knowledge as the crucial element of an understanding of the social world. The social world is a world of theories insofar as it is to a large extent an unknown world for its members (and conversely it is unknown precisely because of the cropping up of theories). So, social actors and theoreticians in general have to understand something quite similar as it is constituted by a universe of socially and culturally generated notions; the object to which this knowledge refers lacks the desirable exteriority. Being the objects of people's concern external and internal at once, we have then, according to Hayek's picture, forces of praxis and poiesis launched simultaneously.

Now, the Hayekian message seems sometimes to be: we have no way to overcome our positional differences, and, accordingly we are unable to attain such an external viewpoint. Yet, Hayek understates an insight of his to the effect that the only method we have available to find our way in this world of uncertainty is to somehow construct such a world in the form of suppositions, anticipations and interpretations of the actions or suppositions of our cohabitants. This insight is better explored by Polanyi. In fact, Polanyi would argue that we have a drive to try to overcome our differences, to transcend our experience. And Popper, even more strongly (and morally?), would contend that we have a duty to invent if we are unable to accomplish the traditional ideal of objectivity. How can we reconcile these premises of our actions? Should we give up one of them?

Let us attempt to understand the components of this world of theories which our social world happens to be. According to the Hayekian view, it is possible to individuate three kinds of theories behind social actors' acts. Firstly, there are local theories which guide ordinary action of actors and which are tantamount to the concrete knowledge available to each according to his position within

the social landscape. Secondly come social theories well advised by a good social philosophy that should have warned the specialists in the matter that the only possible social theories are but a set of ideal types. Thirdly, there are speculative 'vulgar' theories which arise from the synoptical illusion of social actors, an illusion that would mislead them into building up theories of the functioning of the entire social order, but which cannot get rid of their inductive origin and ought to be corrected by properly thought-out social theories.

In this picture, any changes ought to arise just from chance (as adaptive solutions that we stumble upon, or else as mistakes or misunderstandings eventually produced by our local theories), or may be induced by proper social theories on a piecemeal basis. A supplement to this picture, one that tries to retrieve the dismissed synoptical illusion of ours, is suggested by our interpretation of Popper and Polanyi, for they seem implicitly to indicate two alternative sources of change. Popper straightforwardly states the possibility of invention of new worlds from within the very invented tradition of (critical) reason, and Polanyi argues for the possibility of the emergence of a new solution from contrived efforts aiming at it even though an important part of the process might be unspecifiable.

Turning to our question of how to reconcile two basic premises of our action which are, respectively, our particular position within the whole (praxis) and our striving to encompass the whole itself (poiesis), we so far have no better answer than to remain within their very terms. One accepts the indeterminacy not as a failure but as a sign that we have so far succeeded in what our task possibly is.

CONCLUDING CHAPTER

This thesis has identified and commented on three individualist arguments concerning the possibility of social order, three different conceptions of the way in which the social order may result from the interaction of individuals. From the perspective I have taken up here, the different ways in which an individualist conception may be achieved differ in terms of degrees of reductionism and kinds of idealization. Regarding reductionism, I have assumed that any individualistic account of the social order naturally undertakes, although to a variable extent, some reduction of the whole of social order into its individual parts. Hence, the 'one', the 'two', and the 'many'. As for idealization, I have presumed that there is some sense of transcendence involved when we shift from the individuals to the whole, since individuals are seen less in their characteristic 'individuality' and more in their ability to mutually coordinate their actions in some order-enabling way. In this way, my taxonomy has called attention to the different kinds of idealization where individuals may simply coordinate in some order-preserving way, or else adjust their actions in a more stringent equilibrium sense, or coordinate cooperatively or efficiently, or still achieve higher ideals such as justice and well-orderliness through interaction.

It has also been found here that the more reductionist the approach, the 'more' idealized its vision of the social order. So far, so good. Yet, is reductionism convincing? The possibility of reduction has seemed to me to depend on the capacities and motivation with which individuals are endowed in each approach, in particular their presumed ability to abridge the whole community in themselves which turns ultimately on the postulated knowledge possibilities. From an examination of these aspects and relations, I have found reductionism quite unconvincing. If so, how about high ideals? Should we give up them? I have also tried to argue against this skepticism in the last section of Part III. Let me retrace my steps in this thesis and call attention to the structure of the argument I have built up on my way to looking for some answers.

* * *

In the Introductory Chapter, I have tried out a couple of analytical remarks, in an effort to demarcate the set of theories in terms of some traits that would fit into my purpose of discussing reductionism, normative aspects, and epistemological elements, as well as their mutual relations in the theories. In this regard, I have proposed a two-dimensional taxonomy of social action considering different kinds of rationality (parametric, strategic, and evolutionary) and ends of the individuals (public and non-public), corresponding roughly to distinct degrees of reductionism and normative outlooks, which, I have presumed, have a special relation to epistemological elements in the analysis. The classification undertaken should have drawn attention to common characteristics of theories in each cell as well as to the diversity between different cells. In matrix 1, I have also initially characterized the different theories in terms of their belonging to the same perspective of seeing the social order as the result of peculiar individual rational choices. However, with the inclusion of the 'evolutionary' rationality in the modified matrix 2, our reasoning is forced into realizing how enlarged may the conception of 'individual' be as well as the idea that the social order is the result of individual actions and purposes. To see that the subsets of theories in matrix 2 still have much in common and may be seen as distinct interpretations of largely the same tradition, one has to move a long way back into the past.

So, I have tried, in chapter 1, to locate the entire lot of arguments in a common tradition that goes back to the eighteenth century European political thought and that evolves from a discussion over the meaning of the public interest. In that context, the notion of self-interest acquired great significance. Self-interest was proposed as a promising proto-democratic political category that could oppose partisanship in the competition for a suitable basis for the public interest. It was also seen as potentially helping in its capacity to add to the social cement, as in Mandevillian and Smithian arguments for that matter. All this notwithstanding, except for the utilitarian branch of the tradition as denounced by Macaulay, almost every argument redeeming self-interest from its pre-modern exile called attention to its merely vicarious or imputed character in political arguments. In other words, there was nothing self-evident about self-interest nor was it always true that when considered as a natural motivation of a person it necessarily worked for his good, let alone the common good. So, self-interest

reappeared as a **category** shaped in political arguments, whose value in terms of promoting one's or everybody's good was to be established through acceptable meanings conveyed to it by the social environment, and whose primary function was that of being a critical and reflective category.

* * *

Having established this standpoint regarding self-interest, I have proceeded by examining some contemporary theories that stem from the tradition such as those in Parts I, II, and III. My first case, in chapter 2, was a set of theories that typically claims to provide an account of social order in terms of an efficient (and sometimes also just) cooperation among individuals that results from a collective action of self-interested people, or what they call an 'internal solution' to the problem of generating the political order. According to these theories, self-interest may effect the 'bootstrapping' of promoting the public interest. The self-interest of people is here translated into their maximizing behavior in the context of Hobbesian strategic interaction. The perspectives examined assume either complete information, or incomplete information or still endogenous preferences, relying, respectively, on rational choice approaches to individual behavior, institutional design and preferences. I have organized my critical analysis of this literature in a way that contrasted it with the opposite perspective of 'external' solutions which have seemed to me to be more in line with the advice given by the tradition. In the end, internal solutions have appeared to call for external aid of the independent substantive assumption of cooperative behavior or, more generally, to reinforce the impression that external elements to the maximizing utility assumption are needed for cooperation among individuals to arise.

In this direction, I have identified what I have termed an 'identitarian' collective action approach which emphasizes the importance of considering a person as an enlarged entity rather than a bare self, and as possessing constitutive **group** interests or commitments to values and norms that reach beyond their 'selves' and indeed are the condition of the possibility of any maximizing calculation of theirs. Such an enlarged behavioral assumption would render cooperation possible. Two difficulties have stood in the way, according to my analysis. They

refer, on the one hand, to the conceptual confusion between *explanans* and *explananda* where group-interest achievement assumes group-interest motivation, and on the other hand, to the doubtful desirability of this achievement in view of the warnings that come from the tradition about deleterious effects of successful group interest on **public** interest. Is there something **beyond** self-interest that does not belong in the group-interest sphere, or more generally in the altruistic- or, to put it in Kantian terms, 'happiness'- sphere? The claim by accounts discussed in Part II is that there is such a sphere, the moral domain of human motivation. (Note the link between chapter 2 in Part I, and Part II). Indeed, Part II begins by posing a stylized question concerning the narrow limits of what Kant calls the domain of 'generalizability' in contrast to the larger ones of the domain of 'universalizability'.

Before coming to this, however, there are other threads worth pursuing in Part I. To begin with, I have discussed game-theoretical issues in this Part with the intention of outlining some of the theory of rationality behind the public-good theories which spells out how individuals reason in a strategic environment, and, more generally, what is peculiar about a strategic interaction. While the 'public' aspect of rational choices in the sense of my matrix 1 was explicitly discussed in the chapter on order as a public good and had a natural sequence in Part II, their 'strategic' aspect, in turn, was approached in the chapter on game theory. Another important motivation for discussing game theoretical aspects is the underlying assumption in Part I that the notion of equilibrium is an important part to the notion of order (besides its efficiency or justice attributes), and game theory is concerned with producing equilibrium solutions to strategic situations. Given, however, that the environment is strategic in the sense that the individual choosers are aware of their interdependence, epistemic aspects acquire great importance when choosing. In view of this, the equilibrium notions in the standard approach assume at some point a convergence of beliefs among the agents. The best accomplished justification for this convergence assumption I have found within the game-theoretical field has been provided by Bayesianism in its use of the Bayes rule as a learning mechanism and in its adoption of the common priors assumption and the so-called Harsanyi doctrine. These assumptions provide the symmetry conditions for at least **two** distinct individual agents to reason to the equilibrium-profile in an uncertain environment. So, while this account puts stress on the problem of mutual **predictability** and the medium by which

intentions are translated into actions, the attempt to build up a formal reasoning to underscore equilibrium considerations reduces the question of predictability to the seemingly more tractable and less interesting one of 'outguessing'. A series of objections may be raised against the symmetry postulates, and the line I have followed in the thesis has stressed the low philosophical density thereof. My contention is that they take too much for granted, as in the statement that beliefs are coextensive with information where the issue of knowledge in the social interaction is trivialized. In this respect, a better philosophically situated stand is taken up by the approach that I have examined in Part III, and the link between chapter 3 in Part I and Part III is quite direct for that matter.

* * *

Retrieving the link between Parts I and II, we should note that the adoption of the assumption of cooperative behavior on the part of the individuals finds argumentative support in the theories examined in Part II. So, in Part II I have considered a different family of theories stemming from the broad self-interest tradition that produces arguments for the existence of a motivational sphere beyond self-interest, the moral domain; this set of theories illustrates the parametric-public case. The moral sphere is supposed to provide the basis for a well-ordered society, where we would find not only equilibrium and efficiency but also an objective measure of better-orderliness in terms of the achievement of certain ideals of justice.

In the historical chapter 1, I have discussed possible ways of reconciling the self- and the public- interests that one may find in the tradition, and I have mentioned the two theories of laws as edges, and laws of justice, by Hobbes and Locke, respectively. Then, I have tackled the claim by Locke that the laws should also include some positive high ideals and not only a concern about 'bogs and precipices', that is to say, they should also indicate some ends to follow, reflecting a portion of the General Good in each individual. My feeling then was that Locke, like many others including Adam Smith, was well aware of the limits of the self-interested motivation in promoting the General Good and that something else was needed for this to be brought about. Normative theories typically assume the possibility of this 'something else', which they call the moral motivation.

In order to delineate the general traits of these approaches I have written a chapter on Arrow's impossibility theorem which I have interpreted as posing an important question to liberal conceptions of the social order, namely, can preference-autonomy lead to an acceptable notion of social preferences, where this autonomy may involve people's personal tastes as well as values? Or, in other words, can the myriad of interpersonal comparisons people make when forming and expressing their preferences over alternative social states, from their own however enlarged personal standpoint, produce acceptable social states? This I have called Arrow's inductive question concerning the (im)possibility of social order as an aggregate of personal preferences. My attention was drawn to what appeared to me to be the suggestion of a conceptual relation between order and justice, where justice acquired a preeminence, as a condition of possibility of order itself. It was also drawn to what I have taken as a warning as to the difficulty of establishing the justice viewpoint from a 'generalizability' perspective, that is, from a majority of personal viewpoints regarding the social preferences.

Then, in my interpretation, Harsanyi and Rawls' arguments follow those of Arrow's as proposals of construction of an impersonal or impartial standpoint, where the social order is seen as arising from prior well-founded interpersonal comparisons which stem from an agreement on rational principles. Harsanyi finds a support for this standpoint in what he calls our moral (in contrast to personal) preferences and he works it out from a sympathy-like perspective. Rawls poses our sense of justice as providing such a support. Their arguments, although differing in important respects, share the basic common characteristic of approaching the objective standard for interpersonal comparisons (which is the condition of possibility of a well-ordered society) through the thought experiment of one rational individual whose ends have been properly constrained. In a way, either sympathy or the sense of justice may be seen as symmetry devices that enable reason as impersonated in any rational individual to give access to the standpoint of all; in the case of Harsanyi, the so-called Harsanyi doctrine plays a crucial role in the construction of the ideal observer's standpoint. These devices are proposed as elements in the construction of a 'universalizability' perspective, in order to fill the blanks of the inductive construction of the social preferences. (There is an interesting parallel that may be made between the formal argument of the impossibility theorem and the ethical arguments

issued by Harsanyi and Rawls and it is that these latter also deal with the problem of aggregation of individual ends from a formal perspective, where even some substantive aspects of ethical doctrines such as claims about specific ends that individuals or their social arrangements should follow are formally provided. This renders the joint treatment of these theories a much easier and natural task to undertake, where consistency conditions along with moral constraints are only different kinds of formal restrictions to be imposed on one's rational choices over social arrangements.)

A question arises how much objectivity can this new standpoint achieve due to this new set of restrictions? The extent to which intrapersonal deliberation may cross the boundaries between people towards the achievement of an impartial outlook depends much, in these approaches, on the presence of certain knowledge conditions, as I have remarked on in my analysis. This is, in my judgment, where the 'universalizability' argument seems to collapse into the 'generalizability' one, and what I have called the deductive approach seems not to overcome the inductive difficulties in providing conditions for the formation of acceptable social preferences. Reductionism and 'high' ideals go together here: reason as it occurs in **any** rational individual renders one able to design the principles that should regulate the ideal social order, where the demands of stability meet those of justice. However, the objective knowledge that is needed to the achievement of the **same** set of principles by any rational individual is reduced into a less infallible theoretical knowledge.

Certain nuances are worth remarking on in this Part. Generally speaking, the transcendence aspect involved in any individualistic account of social order is here thought of in terms of the access individual reason conveys to an ideal community of individuals, through principles of universal validity which it may reach, understand and articulate. In the rule-utilitarian version there is but one **rational good** to be pursued, and this comes out as the result of an individual's properly constrained hypothetical reasoning. Hence, an ideal social order is that which is ordered according to principles which result from a rational individual's understanding of what the one rational good is. This view is unacceptable to Rawls, who launches a very important criticism to utilitarianism on the appealing ground of the **incommensurability** of the goods, or of conceptions thereof, in the society. This is a 'fact' which our ideal conception

of social order should address somehow. However, Rawls' view of social order is still aggregative, for he also looks for the one rational standpoint whereby a set of principles would unfold, although with a view that there is no trade between those goods and therefore that we should think in terms of a hierarchy or a serial order of common regulative ends which we all share, regardless of our differing conceptions of the good. In the earlier Rawlsian argument of A Theory of Justice, room is given to an incommensurability between goods as pursued by rational individuals, the emphasis being placed on the possibility that rational individuals have a common regulative interest in abiding by public principles which in turn gains support from their sense of justice. In the later Rawlsian argument one does not attain universal principles only by means of ignoring one's particular position in the social arrangement and relying on one's sense of justice. This time, the achievement of public principles is also constrained by what the many radically incommensurable views of the good might all share, relying on a notion of public reason and the demands of the 'Reasonable'. and hence getting closer to a view of 'justice as order' rather than 'order as justice'.

Moving from the rule-utilitarian to the later Rawlsian arguments it becomes increasingly difficult to reduce the social order into one individual's rational endeavors, as well as to harmonize the different ideals to which the order so conceived should be conducive. Nonetheless, it seems to me that there is something strikingly similar in these approaches to the effect that the 'representation fact' is conceived unproblematically, or non-philosophically. The idealized representation of the social order in terms of the interaction of rational individuals take these latter as only capable, in the processes of deciding on what principles to organize the basis of their interaction, of following a proper method of reasoning and obtaining the relevant information to this end (leaving aside the psychological assumptions of sympathy and sense of justice that warrant the ideal agreement).

* * *

A different view of individual transcendence towards the construction of a social order is displayed in the third approach that I have examined in Part III. It derives much of its vision of the social world from the eighteenth century tradition that I have discussed in chapter 1,

indeed it may be argued that this perspective is a legitimate heir of that tradition as it has been interpreted in the present dissertation. In the chapter on self-interest I have pointed out certain quite obvious limitations of this behavioral assumption, in terms of its restricted use either in explanatory or in normative arguments due basically to the fact that, to take a metaphor suggested in the title to that chapter, self-interest does not possess an absolute **value** but only a relative one. In other words, its meaning is relative to one's circumstances and to the knowledge open to him, relative to more basic and less desirable motivations of human action, relative to political argument and, more generally, relative to opinions. In a nut-shell, self-interest is highly sensitive to the **restrictions**, be they moral, political or epistemic, to an extent that renders it almost empty when it is asked to yield an explanation or a normative horizon to human interaction in its terms. So then, this basic idea is captured by the third set of thoughts that I have analyzed in the thesis under the heading of 'interacting individuals'.[1]

This is an approach that clearly emphasizes the role of restrictions on social action in the construction of our social world in a way that downplays the importance of the plain behavioral assumption of self-interest and, in general, of individual rationality. What is more, it also proposes that the motivation of self-interest itself is something that is given meaning by the socially evolving restrictions. However, as it might be supposed, it does not involve the thorough rejection of individualism as in a holistic approach where the social world is explained in terms of some prior social entities and individual action is in the same way determined by these pre-existing social forms. The vision here is of a complex relationship between the irreducible domains of the social and the individual. I have tried to get closer to this vision and to offer its anatomy.

What comes out is an intermediary position between canonical holist and individualist approaches to social order, which I have labeled the 'interacting individuals' variety. In contrast to the individualistic varieties that I have gone through in the present thesis, this outlook offers

[1] At the precise time I wrote the first version of this chapter and submitted it to my supervisor in 1995, I did not know the article by Alan Kirman, "Economies with Interacting Agents". After knowing it, I have decided to maintain my title because the distinction between individuals and agents has seemed to me to be significant enough. I have been most interested in understanding interacting individuals as peculiar carriers of knowledge rather than agents.

a view of order not as a rational choice of either maximizing agents or (also) moral persons but as an emerging unintended effect of the interaction of 'interacting individuals' over time. Before summing up the main steps I have taken towards that end, let me recall that the Hayekian vision of society as a complex interactive system is not new as it comes from the tradition I have examined in chapter 1, in particular from Mandeville and the Scottish enlightenment authors whose ideas Hayek explicitly interpreted and elaborated upon. It is also worth remarking that Hayek's vision bears important similarities to those of British philosopher Edmund Burke and, more recently, British historian Michael Oakeshott and Austrian thinkers such as Michael Polanyi and Karl Popper. With the exception of Oakeshott who takes a clearly anti-rationalistic stand, these authors claim with varying intensities a limited role for reason (as it is available in each individual) to account for the social world, and oppose either reason's 'conceit' or 'immodesty'. I have assumed here that Scottish enlightenment can provide an interesting defence of (a modest form of) rationalism. However, before coming to that let me resume my interpretation of Hayek's concerning his conception of the social world.

Chapter 7 on Hayek is divided into 7 sections which pursue the following scheme. The first section provides the Hayekian general outlook concerning the social order as an emerging effect of individual actions, where the macro-level of the social order appears as a surprising effect of the actions and interactions of individuals otherwise motivated and radically ignorant. The micro-level of the individuals is looked into in sections 2, 3, and 4. In section 3, in particular, the Hayekian 'true' individualism is displayed via the image of interacting individuals, that is, individuals that only exist and make sense in society, where reductionism is rejected and the importance of an intermediary or intersubjective space is affirmed that takes the form of 'rules'. These are proposed as a truly ontological dimension that does not, however, collapse into a collectivistic, let alone, individualistic essentialism. The nature of rules is epistemic; they are **social** forms of knowledge that have their origin in the interaction of the individuals over time and whose role is to complement the limited reach of individual rationality to go about the social intercourse. These rules, in other words, cannot be entirely known in the individuals' consciousness.

What kind of knowledge is this? In contrast to the views displayed in the previous Parts of the thesis, this knowledge is not understood as unambiguous information and methods of reasoning, it is instead a more problematic notion of knowledge that we have to deal with here. Rules are proposed in a *Verstehen* sense, as internal-external viewpoints, that is, objective norms subjectively or individually consumed and being produced by the same act whereby they are consumed. Since the individual does not have an access through reason to the rules he follows or should follow as they cannot be entirely known in his consciousness, he can only get the rules from his particular position in the social world. Objectivity in this way has an irreducible personal coefficient, to borrow Polanyi's expression. However, as the individual catches the rule on his way to understanding or acting, he adds to the general stock of knowledge to be mastered by his fellow actors, thus his subjective acts have an objective however unintended extension.

From this viewpoint, we have the consequence that any knowledge is only practical knowledge, that is, limited to one's particular position in the world and interpretation apparatus, although being objective in the sense that it extends far beyond the individual sphere. In this sense, the individuals' positions are not interchangeable and the possibility for people to coordinate on their reciprocal plans derives from analogy rather than symmetry considerations. In section 4, the double aspect of knowledge, the objective and the subjective, is discussed in depth, and we see that every aspect that concerns the individual action in the social world is rule-laden, comprising his perceptions, experiences, purposes and decisions.

It is also shown in that section that the way the individuals understand and apply the rules, which they do in virtue of their ignorance, is witness, in a non-intuitive manner, to their additive or creative capacity. Because they cannot have before them a clear-cut idea of the rules they follow as they cannot articulate them all in their minds, the rules are subjected to the individual's interpretative apparatus: the understanding of the situation in which they find themselves as well as the comprehension of the rules applicable and use thereof are submitted to the personal coefficient. I believe that this vision permits a fresh interpretation of the famous dictum by Adam Ferguson, that the social world is the product of human action not of design, though I can only give a partial idea of what I have in mind at this point. So, the fact

that not every rule that we follow can be articulated in a conscious manner **does not rule out** our creative capacity and does not make of us wholly determined creatures. Further on I will reintroduce this dictum in connection with Polanyi and Popper's contribution.

In sections 5 and 6, I have taken the path back from the micro- to the macro-level on the way to devising the Hayekian idea of order, the spontaneous order. A series of elements comes into the analysis to yield the image of an autonomous or complex and changing structure. Then, I have captured the tension that dwells in the heart of Hayek's idea of order, as this is connected with preservation and change, by distinguishing the 'evolutionist' from the 'evolutionary' aspects of Hayek's thought. The evolutionary Hayek is at odds with the evolutionist Hayek as the epistemic (evolutionary) perspective sets limits to the normative (evolutionist) viewpoint.

In section 7 a way out is proposed that draws on Popper and Polanyi's ideas. The startling thing about this route is that it suggests a possible reconciliation between the epistemological aspects of Hayek's analysis and personal deliberation and intervention, in a way that enables an original relationship between ignorance and design, between, that is, our epistemic condition and our inclination to actively change our world. This gives a new flavor to Ferguson's dictum that order is the product of action not of will, in that while the constructivist alternative is rejected on the grounds of our precarious private stock of reason, to put it in Burke's words, this action is not completely determined by the blind forces of 'natural' selection. Between full-blown intentional action and 'naturally' determined action there seems to exist a third alternative, an intriguing third bank of the river, as in the image of João Guimarães Rosa. This possibility seems to come out from Michael Polanyi's discussion of personal knowledge, where, I believe, it is possible to catch a glimpse of a new micro (corresponding to complex structures such as society, scientific discovery and language) in the notion of a **passive action**, which has not the full strength of intentional action nor the weakness of passivity and which is a **contrived** effort that causes an **emergence**. This also comes out from Popperian imaginative reason, where **circumstances** are approached in a 'constructivist' way instead of in the external and rather natural manner often present in so-called evolutionary approaches.

In relation to previous Parts of the thesis, I have to say that Part III has a direct link with Part I, especially the chapter on game theory, in terms of its epistemological discussion and the limits it imposes on symmetry considerations and the role of individual rationality. Moreover, its dynamic view of order does not fit well into standard game-theory's static stand in this regard. Of course, the form that the discussion has taken is clearly different here, as I have discussed the epistemological aspects of Hayek's **social philosophy** while in the chapter on game-theory I have tried to spell out its epistemology from within a more **theoretical** discussion (though I have come to understand that Hayekian social theory can only be a social philosophy because it is reflective and suspicious of its own findings!). However, in general terms, in both cases the effort has been made to characterize their views of interaction and, then, the reduction of social interaction into smaller parts, namely, strategic rationality of maximizing agents in one case, or limited rationality of individual rule-followers in the other.

As for the relation to Part II, a series of aspects may be raised. With respect to utilitarianism, Hayek is clearly a critic of its disregarding ignorance and the corollary impossibility of **calculating** the full consequences of one's acts in an extended society. The Archimedian point that might be built up on the basis of such a rationalistic conceit is therefore dismissed by Hayek. Regarding Rawls, what this epistemological Hayekian discussion objects is to the impossibility of fully **articulating** the rules that we follow while acting, comprising our sense of justice, which renders the enterprise of justification of general principles always partial. Reason, with a capital R, is not always fully and equally available to all, notes Hayek. Considering either the rule-utilitarian argument or that of Rawls' , Hayek's view of knowledge is quite distinct in that for him the possibility of knowledge in the social world is itself interestingly problematic and is not solved by reducing knowledge to non-disputable data and correct methods of reasoning that may be fully and equally available to any individual. Knowledge is the moving stuff out of which the social world is made, for although it is condensed in these social forms such as the system of rules, it is always irreducibly intersubjective and undergoing change as long as these rules are individually consumed and circumstances change. It may be said that the cases that I have examined in Part II also postulate a social dimension in each individual action, as Harsanyi postulates the social nature of our preferences and Rawls recalls the compelling demands on our acts of our sense of

justice and of what he calls the 'reasonable'. The presence of this dimension in the individual domain opens the possibility for the morality of our individual actions.

In Hayek, in contrast, although the social dimension is said to be present in every act of ours and even our intuitions to have a social basis, this presence does not constitute a reliable basis for the morality of our acts as the multi-level rules we follow cannot be fully articulated at the individual level, and our knowledge of these can only be partial and relative to our position in the social world. Our actions as ordinary actors are always mediated by theories or hypotheses, efforts of understanding through classification (in the same ways as extraordinary actors such as social theoreticians do). The only truly impersonal or impartial standpoint is that offered by the evolving structure represented by society, that is, this ontological dynamic system that cannot be understood in the sense of the articulable knowledge available to one individual. Still, in terms of the normative consequences of such a picture, while utilitarianism claims the existence of what Rawls has called the one rational good, and Rawls himself claims that there is incommensurability of goods but not of rights and that, as it were, there can only be the common meta-good of toleration, Hayek seems to claim the incommensurability of persons which may be derived from his criticism of essentialistic individualism. Is that all there is to say about normative horizons?

In this connection, the integration of Hayekian epistemology with pieces of Popper and Polanyi's thinking has been made with the purpose of experimenting the possibility of reconciliation between Hayekian epistemology and design, something Hayek and his more conservative colleagues have always rejected quite boldly. The Hayekian warning that design would unleash perverse consequences has always sounded a bit odd in view of his very convincing epistemological arguments that we cannot fully know the consequences of our acts. The Hayekian contention that evolution pursues some end has equally sounded awkward in view of the same quite illuminating epistemological findings of his. So, he cannot as safely state that everything tends to the better or the worse for the same very reasons provided by the epistemological argument.

So far so good. Still, I keep wondering, is it possible to make more out of this? Could we possibly hope for **hope** and yet accept our epistemic condition? I have argued here that Popper and Polanyi's ideas provide a basis for such a hope. Popper proposes a modest rationalism that understands itself as a tradition, having born from the invention of a tradition, namely, the tradition of criticism (a view that bears similarity to the eighteenth century outlook regarding self-interest as a reflective category). Polanyi identifies the passive action of an actor endowed with personal knowledge whose roots are twofold, conscious and unconscious, and where the conscious side pushes him forward in a way he is unable to describe fully. He locates the intentional action in between its inarticulable backward condition and its thoroughly unpredicted consequences, preserving however its sense of deliberation and creativity, in what seems to me to be a *poietical* construction of the social world

* * *

Having said that, I should say a final word about the kind of integration of these theories that is possible within the limits of my undertaking here. I want to call attention to the fact that this integration as it exists in the thesis is only a **meta-theoretical** one, where a set of theories have been examined in view of their belonging to a common tradition of ideas, which I have characterized as such, and their distinct answers to some salient questions that I have suggested pertain to this tradition. A **theoretical** integration where better argued propositions of one particular theory would be imported into another in order to overcome the latter's difficulties was not my purpose here.

* * *

As an afterword, the following diagram[2] may give a synoptic (though perhaps too schematic) idea of my three ideal types in terms of their idealized views of social order, their assumptions concerning the motivation of the individuals, the degrees of reductionism involved, and the knowledge possibilities open to the actors:

[2] Note that the elements in the cells are just the focal (and not the exclusive) characteristics of each type.

IDEAL TYPE	ORDER AS	INDIVIDUALS AS	REDUCTION	KNOWLEDGE POSSIBILITIES
SELF-INTEREST	Efficiency/ Equilibrium	Utility- maximizers	Otherness (the two)	Information and method[3]
AUTONOMY	Justice	Moral persons	Uniformity (the one)	Information and method[4]
INTERACTING INDIVIDUALS	Change	Rule- followers	Complexity (the many)	articulable and inarticulable rules

[3] This means 'unambiguous' information and 'correct' methods of reasoning.

[4] See footnote 3.

References

ARROW,K.,(1951,1963), Social Choice and Individual Values, New York: John Wiley and Sons.

ARROW,K.,(1984), Social Choice and Justice (Collected Papers of Kenneth J. Arrow, Vol.1), Belknap Press.

ARROW,K. ET AL. (1996), The Rational Foundations of Economic Behaviour, London: MacMillan Press Ltd.

AUMANN,R., (1976), "Agreeing to Disagree", The Annals of Statistics, Vol.4,No.6,1236-1239.

AUMANN,R., (1987), "Correlated Equilibrium as an expression of Bayesian rationality", Econometrica, vol.55, no 1, pp.1-18.

AXELROD,R., (1984), The Evolution of Cooperation, New York: Basic Books.

BASU,K., (1990), "On the Non-Existence of a Rationality Definition for Extensive Games", International Journal of Game Theory 19, 33-44.

BAYNES,K.,(1992), The Normative Grounds of Social Criticism - Kant, Rawls, and Habermas, New York: State University of New York Press.

BEINER,R. AND BOOTH,W.J., (1993), Kant and Political Philosophy - The Contemporary Legacy, New Haven and London: Yale University Press.

BICCHIERI,C., (1993), Rationality and Coordination, Cambridge: Cambridge University Press.

BINMORE,K., (1993), "De-Bayesing Game Theory", IN: Binmore, Kirman and Tany (eds.), Frontiers of Game Theory.

BINMORE,K., (1994), Playing Fair: game theory and the social contract, Cambridge, Mass.: The MIT Press.

BINMORE,K., KIRMAN,A., TANI,P., (eds.), (1993), Frontiers of Game Theory, Massachusetts Institute of Technology.

BROOME,J.(1991), Weighing Goods, Cambridge: Basil Blackwell.

BURKE, E., (1790,1955), Reflections on the Revolution in France. New York: The Bobbs-Merril Company, Inc.

CALDWELL, B.,(1994), "Hayek's Scientific Subjectivism", Economics and Philosophy, vol x, pp.305-313.

CROWLEY,B.L., (1987), The Self, the Individual and the Community - Liberalism in the Political Thought of F.A.Hayek and Sydney and Beatrice Webb, Oxford: Clarendon Press.

CULLITY,G., (1995), "Moral Free Riding", Philosophy and Public Affairs, Winter 1995, vol.24, no.1 .

DE JASAY,A., (1989), Social Contract, Free ride - a study of the public goods problem, Oxford: Clarendon Press.

DUPUY,J.-P., (1989), "Common knowledge, Common sense", Theory and Decision 27, 37-62.

DUPUY,J.-P., (s.d.), "Temps et Rationalité", mimeo.

EATWELL,J., MILGATE,M., AND NEWMAN,P., (eds.), (1987), The New Palgrave - a dictionary of economics, London: The Macmillan Press Limited.

EATWELL,J., MILGATE,M., AND NEWMAN,P., (eds.), (1990), The New Palgrave - Utility and Probability, London: The Macmillan Press Limited

ELSTER,J.(1979), ,Ulysses and the Sirens - studies in rationality and irrationality, Cambridge: Cambridge University Press.

ELSTER,J.,(1982), "Sour Grapes - utilitarianism and the genesis of wants", IN: A.SEN and B.WILLIAMS (eds.), Utilitarianism and Beyond.

ELSTER,J. (1985), "Rationality, Morality, and Collective Action", Ethics,96 (October 1985):136-155.

ELSTER,J. (1989), The Cement of Society: a study of social order, Cambridge: Cambridge University Press.

ELSTER,J., (1990), "Selfishness and Altruism", IN: J. MANSBRIDGE (ed.), Beyond Self-Interest.

ELSTER,J. AND HYLLAND,A.,(eds.) (1986), Foundations of Social Choice Theory, Cambridge: Cambridge University Press.

ELSTER,J. AND ROEMER,J.(eds.), (1991), Interpersonal Comparisons of Well-Being, Cambridge: Cambridge University Press.

ESPADA, J.C., (1996), Social Citizenship Rights: A Critique of F.A.Hayek and Raymond Plant, London and New York: Macmillan/St Martin's Press.

FERGUSON,A., History of Civil Society, London: Forbes.

FOXLEY,A., McPHERSON,M.S. AND O'DONNEL,G.(eds.), (1986), Development, Democracy and the Art of Trespassing: essays in honor of Albert O. Hirschman, Notre Dame, Indiana: University of Notre Dame Press.

GAUTHIER,D., (1986) Morals by Agreement, Oxford: Clarendon Press.

GOOD,I.J., (1990), "Subjective Probability", IN: J.EATWELL, M.MILGATE AND P.NEWMAN (eds.), The New Palgrave - Utility and Probability.

GOUGH,J.W., (1957), The Social Contract - A critical Study of its Development, Oxford: Clarendon Press.

GUNN,J.A.W., (1969), Politics and the Public Interest in the Seventeenth Century, London: Routledge & Kegan Paul.

HALEVY,E.,(1972), The Growth of Philosophical Radicalism, London: Faber & Faber.

HAMMOND,P.,(1991a), "Interpersonal Comparisons of Utility: Why and how they are and should be made", IN: J.ELSTER and J.ROEMER (eds.), Interpersonal Comparisons of Well-Being.

HAMMOND,P.,(1991b), "Harsanyi's Utilitarian Theorem: A Simpler Proof and Some Ethical Connotations", EUI Working Paper ECO no. 91/32, European University Institute, Florence, Economics Department.

HAMPTON,J.,(1989), Hobbes and The Social Contract Tradition, Cambridge: Cambridge University Press.

HAMPTON,J.,(1994), "The Failure of Expected-Utility Theory as a Theory of Reason", Economics and Philosophy, 19 (1994), 195-242.

HARDIN,R., (1971), "Collective Action as an Agreeable n-Prisoner's Dilemma", Behavioral Science 16, 472-481.

HARDIN,R., (1984), "Difficulties in the notion of economic rationality", Social Science Information,23,3 (1984).

HARDIN,R., (1988), Morality within the limits of reason, Chicago: The University of Chicago Press.

HARDIN,R., (1991), "Hobbesian Political Order", Political Theory, Vol.19 No.2, pp.156-180.

HARDIN,R., (1992), "Determinacy and Rational Choice", IN: R.SELTEN,(ed.), Rational Interaction.

HARDIN,R., (1995), One for All: the logic of group conflict, Princeton: Princeton University Press.

HARSANYI,J.C.,(1958,1980), "Ethics in Terms of Hypothetical Imperatives", IN: J.C.HARSANYI, Essays on Ethics, Social Behavior, and Scientific Explanation.

HARSANYI,J.C., (1967/1968), "Games with Incomplete Information Played by 'Bayesian' Players". Parts I, II and III, Management Science, 14, 159-82, 320-34, 486-502.

HARSANYI,J.C., (1973,1980), "Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory", IN: J.C.Harsanyi, Essays on Ethics, Social Behavior, and Scientific Explanation.

HARSANYI,J.C., (1973,1982), Papers in Game Theory, Dordrecht: D. Reidel Publishing Company.

HARSANYI,J.C., (1980), Essays on Ethics, Social Behavior, and Scientific Explanation, London: Reidel.

HARSANYI,J.C., (1982), "Morality and the theory of rational behaviour", IN: A. Sen and B. Williams, Utilitarianism and Beyond.

HARSANYI,J.C., AND SELTEN,R., (1988), A General Theory of Equilibrium Selection in Games, Cambridge: MIT Press.

HARSANYI,J.C., (1992), "Utilities, Preferences and Substantive Goods", The United Nations University WIDER, Working Papers no. 101, December 1992.

HAYEK,F.A., (1949), Individualism and Economic Order (IEO), London: Routledge & Kegan Paul Ltd.

HAYEK,F.A., (1952), The Sensory Order, The University of Chicago Press.

HAYEK,F.A., (1964), The Counter-revolution of Science - studies on the abuse of reason (CR), New York: Free Press

HAYEK,F.A., (1967), Studies in Philosophy, Politics and Economics (S), London: Routledge & Kegan Paul.

HAYEK,F.A., (1976), Law, Legislation and Liberty, vol.2 The Mirage of Social Justice, London: Routledge & Kegan Paul.

HAYEK,F.A., (1978) New Studies in Philosophy, Politics, Economics and the History of Ideas (NS), London: Routledge and kegan Paul.

HAYEK,F.A., (1983), Knowledge, Evolution and Society, London: The Adam Smith Institute.

HAYEK,F.A., (1988), "Between Instinct and Reason", IN: The Collected Works of Friedrich August Hayek - vol.1; The Fatal Conceit -the Errors of Socialism, London: Routledge.

HEAP,S.P.H. & VAROUFAKIS,Y. (1995), Game Theory - a critical introduction, London and New York:Routledge.

HECHTER,M., (1992), "The Insufficiency of Game Theory for the Resolution of Real-World Collective Action Problems", Rationality and Society, Vol.4 No.1, January, pp.33-40.

HECKATHORN,D.D., (1989), "Collective Action and the Second-Order Free-Rider Problem", Rationality and Society, Vol.1 No.1, pp. 78-100.

HEILBRONER,R., (1973), "The Paradox of Progress: Decline and Decay in the Wealth of Nations", Journal of the History of Ideas (April-June 1973), pp.242-262.

HIRSCHMAN,A., (1977), The Passions and the Interests, Princeton: Princeton University Press.

HIRSCHMAN,A., (1986), "Self-Interest: From Euphemism to Tautology", IN: A.O.Hirschman, Rival Views of Market Society and Other Essays, New York: Viking.

HOBBES,T., (1968), Leviathan, Harmondsworth: Penguin.

HOLMES,S., (1990), "The Secret History of Self-Interest", IN: J.MANSBRIDGE, (ed.), Beyond Self-Interest.

HOOK,S.(ed.), (1967), Human Values and Economic Policy, New York: New York University Press.

HORNE,T.A., (1978), The Social Thought of Bernard de Mandeville: virtue and commerce in early eighteenth century England, London: The Macmillan Press.

HUME, D., (1985), A Treatise of Human Nature, ed. L.A.Selby-Bigge, Oxford: Clarendon Press.

HUME, D., (1994), Political Writings, ed. S.D.Warner and D.W.Livingston, Indianapolis: Hackett.

HYLLAND, A., (1986), "The purpose and significance of social choice theory: some general remarks and an application to the 'Lady Chatterley problem'", IN: J.ELSTER AND A.HYLLAND, Foundations of Social Choice Theory.

JEVONS,W.S., (1965), The Theory of Political Economy, New York: Reprints of Economic Classics, A.M.Kelley.

JOHNSON, JAMES, (1993), "Is Talk Really Cheap? Prompting Conversation Between Critical Theory and Rational Choice", American Political Science Review, Vol.87, No.1, March.

KANT, I., (1978), The Critique of Practical Reason and other ethical treatises, Chicago: Encyclopaedia Brittanica.

KELSEY, D. AND QUIGGIN, J., (1989), "Behind the Veil: A Survey of Theories of Choice under Ignorance and Uncertainty", Working Paper no. 183, Working Papers in Economics and Econometrics, The Australian National University.

KELSEY, D., (1988), "A Survey of Ignorance or Alternatives to Subjective Expected Utility", Economic Theory Discussion Paper no.129, University of Cambridge, August 1988.

KEYNES,J.M., (1973), The General Theory of Employment, Interest, and Money, Vol.7 of The Collected Writings of John Maynard Keynes, London: Macmillan and Cambridge University Press.

KIRMAN,A.P., (1992), "Whom or What Does the Representative Individual Represent?", Journal of Economic Perspectives, Vol.6, No.2, Spring 1992, pp.117-136.

KIRMAN,A.P.,(1995), "Economies with Interacting Agents", Discussion Paper No.A-500, Rheinische Friedrich-Wilhelms-Universität Bonn.

KNIGHT,J., (1992), Institutions and Social Conflict, Cambridge: Cambridge University Press

KORS AND KORSHIN (eds.), (1987), Anticipations of the Enlightenment in England, France, and Germany, Philadelphia: University of Pennsylvania Press.

KREPS,D., (1990), A Course in Microeconomic Theory, New York: Harvester Wheatsheaf.

KREPS,DAVID M., (1987), "Nash Equilibrium", The New Palgrave - a dictionary of economics, volume 3.

KUKATHAS,C., (1989), Hayek and Modern Liberalism, Oxford: Clarendon Press.

KUKATHAS,C. AND PETTIT,P., (1990), Rawls - A Theory of Justice and its Critics, Polity Press.

LARMORE,C. (1987), Patterns of Moral Complexity, Cambridge: Cambridge University Press.

LASLETT,P., (1988), "Introduction", IN: J.LOCKE, Two Treatises of Government, Cambridge: Cambridge University Press

LEONARD,R.J., (1995), "From Parlor Games to Social Science: von Neumann, Morgenstern, and the Creation of Game Theory 1928-1944", Journal of Economic Literature, vol.xxxiii (June 1995), pp. 730-761.

LESSA,C.A.,(1993), "Ação Coletiva e Contratualismo: Interpretação ou Demonstração?".Nova Economia, vol.xxxi, n.1, pp.21-32.Belo Horizonte, Brasil.

LESSA,C.A., (1995), "Racionalidade e Contingência: O Caso da Ação Coletiva", Arché, ano IV, no.10, 1995.Rio de Janeiro, Brasil.

LEWIN,R., (1993), Complexity - life at the edge of chaos, London:Phoenix.

LEWIS,D. (1969), Convention: A Philosophical Study, Cambridge, Mass.: Harvard University Press.

LEWONTIN, R.C., (1968), "The Concept of Evolution", IN: International Encyclopedia of the Social Sciences, Macmillan & Free Press, pp 202-209.

LIVELY,J. AND REES,J., (eds.), (1978), Utilitarian Logic and Politics, Oxford: Clarendon Press.

LOCKE,J., (1988), Two Treatises of Government, ed. P.Laslett, Cambridge: Cambridge University Press.

LUCE,R.D. AND RAIFFA,H., (1957,1964), Games and Decisions - Introduction and Critical Survey, New York: John Wiley & Sons, Inc.

LUKES, S.(1973), Individualism. Oxford: Basil Blackwell.

MACAULAY,T.B., (1978), "Mill's Essay on Government: Utilitarian Logic and Politics", IN: LIVELY and RESS (eds.) Utilitarian Logic and Politics.

MANSBRIDGE,J., (ed.), (1990), Beyond Self-Interest, Chicago: The University of Chicago Press.

MARX,K., (1977), Capital - A Critique of Political Economy, vol.I, London: Lawrence & Wishart.

MAYNARD SMITH,J., (1982), Evolution and the Theory of Games, Cambridge: Cambridge University Press.

MILLER, D.,(ed.)(1987), A Pocket Popper, Fontana Press.

MIROWSKI,P., (1992), "What were von Neumann and Morgenstern Trying to Accomplish?", IN: E.R. WEINTRAUB (ed), Toward a History of Game Theory.

MOORE,O.K., AND ANDERSON,A.R., (1962), "Some Puzzling Aspects of Social Interaction", The Review of Metaphysics, Vol. XV, No.3, Issue No.59.

MORRIS,S., (1994), "The Common Prior Assumption in Economic Theory", Economics and Philosophy, 11, 227-253.

MULHALL,S. AND SWIFT,A., (1992), Liberals and Communitarians, Oxford:Blackwell.

NASH,J., (1951), "Non-cooperative games", Annals of Mathematics 54, 286-95.

OAKESHOTT,M., (1974), Rationalism in Politics and other essays, London: Methuen & Co Ltd.

OAKESHOTT,M., (1974), "Rationalism in Politics", Rationalism in Politics and other essays.

OAKESHOTT,M., (1974), "Rational Conduct", Rationalism in Politics and other essays.

OAKESHOTT,M., (1974), "Political Education", IN: Rationalism in Politics.

OLSON,M., (1971), The Logic of Collective Action - Public goods and the Theory of Groups, Harvard University Press.

OLSON,M., (1992), "Foreword", IN: T. SANDLER, Collective Action - theory and applications.

ORLÉAN,A., (1989), "Mimetic Contagion and Speculative Bubbles", Theory and Decision, 27 (1989), 63-92.

OUTWAITHE,W., (1975), Understanding Social Life: the method called Verstehen, London: Allen & Unwin.

PIZZORNO,A., (1986), "Some Other Kinds of Otherness: A Critique of 'Rational Choice' Theories", IN:A.FOXLEY, M.S.MCPHERSON and G.O'DONNEL(eds.), Development, Democracy and the Art of Trespassing: essays in honor of Albert O. Hirschman.

POLANYI, M., (1946,1966), Science, Faith and Society (SFS), Chicago: The University of Chicago Press.

POLANYI,M., (1958), Personal Knowledge - Towards a Post-Critical Philosophy (PK), London: Routledge & Kegan Paul.

POPPER,K., (1945,1987), "The Defence of Rationalism". IN: D.MILLER, A Pocket Popper.

POPPER, K., (1949,1987), "Towards a Rational Theory of Tradition". IN: D.MILLER, A Pocket Popper.

POPPER,K.R., (1974), Objective Knowledge - an evolutionary approach, Oxford: Clarendon Press.

RAWLS,J., (1971), A Theory of Justice (TJ), Cambridge, Mass.: Harvard University Press.

RAWLS,J., (1980), "Kantian Constructivism in Moral Theory" (KC), The Journal of Philosophy LXXVII, No.9, September 1980.

RAWLS,J., (1982), "Social Unity and Primary Goods" (SU), IN: A.SEN and B.WILLIAMS (eds.), Utilitarianism and Beyond.

RAWLS,J., (1985), "Justice as Fairness: Political not Metaphysical" (PNM), Philosophy and Public Affairs, vol.14, no.3.

RAWLS,J., (1987), "The Idea of an Overlapping Consensus" (IOC), Oxford Journal of Legal Studies, vol.7, no.1.

RAWLS,J., (1988), "The Priority of Right and Ideas of the Good" (PRIG), Philosophy and Public Affairs, volume 17, number 4 (Fall 1988).

RAWLS,J., (1993a), Political Liberalism (PL), New York: Columbia University Press.

RAWLS,J., (1993b), "Themes in Kant's Moral Philosophy", IN: R. BEINER and W.J.BOOTH, Kant and Political Philosophy - The Contemporary Legacy.

ROSA,J.G., (1962), "A Terceira Margem do Rio", IN: Primeiras Estórias, Rio de Janeiro: Livraria José Olympio Editora.

ROSENBERG,N., (1965), "Adam Smith on the Division of Labor: Two Views or One?", Economica 32 (May 1965), pp.127-139.

SAMUELSON,P., (1967), "Arrow's Mathematical Politics",IN:S.HOOK (ed.), Human Values and Economic Policy.

SANDLER,T., (1992), Collective Action - theory and applications, New York: Harvester Wheatsheaf.

SANTOS,W.G., (1990), "Economia e Ignorancia", Nova Economia, Belo Horizonte, v.1, n.1, novembro 1990.

SAVAGE,L., (1954), The Foundations of Statistics, New York: Wiley.

SCHARPF,F.W., (1990), "Games Real Actors Could Play - the problem of mutual predictibility", Rationality and Society, vol.2 No.4, pp.471-494.

SCHUTZ,A., (1967) The Phenomenology of the Social World, ed. by G. Walsh and F. Lehnert, Evanston: Northwestern University Press.

SCHUTZ,A., (1975), On Phenomenology and Social Relations, ed. by H.R.Wagner, Chicago: University of Chicago Press.

SELTEN,R., (1978), "The Chain Store Paradox", Theory and Decision 9, pp. 127-159.

SELTEN,R., (ed.), (1992), Rational Interaction: essays in honor of J.C.Harsanyi, Berlin: Springer.

SEN,A.,(1970) "The Impossibility of a Paretian Liberal", Journal of Political Economy, 78 (January/February 1970), 152-7.

SEN,A.,(1985a), "Social Choice and Justice: A Review Article", The Journal of Economic Literature, vol.XXIII, no.4, December 1985.

SEN,A., (1985b) "Goals, Commitment, and Identity", Journal of Law, Economics and Organization, vol.1 no.2 (Fall 1985).

SEN,A. AND WILLIAMS,B., (eds.),(1982), Utilitarianism and Beyond, Cambridge: Cambridge University Press.

SMITH,A., (1976), The Theory of Moral Sentiments, Raphael,D.D.(ed.) and Macfie,A.L.(ed.), Oxford: Clarendon Press.

SMITH,A., (1976), An Inquiry into the Nature and Causes of the Wealth of Nations, Oxford: Clarendon Press.

SONG,H.H., (1995), "Adam Smith as an early pioneer of Institutional Individualism", History of Political Economy, vol.27,n0.3 (Fall 1995).

TAYLOR,M., (1987), The Possibility of Cooperation, Cambridge: Cambridge University Press

TAYLOR,M., (1976), Anarchy and Cooperation, London: Wiley.

TSEBELIS,G., (1990), Nested Games: rational choice in comparative politics, Berkeley: University of California Press.

TULLOCK,G., (1992), "Games and Preference", Rationality and Society, Vol.4 No 1, January, pp.24-32.

TURNER,S., (1994), The Social Theory of Practices - tradition, tacit knowledge and presuppositions, Polity Press.

UDEHN,L., (1993), "Twenty-five years with The Logic of Collective Action", Acta Sociologica, vol.36, 239-261.

VAN KLEY,D., (1987), "Pierre Nicole, Jansenism, and the Morality of Enlightened Self-Interest", IN: KORS and KORSHIN (eds.), Anticipations of the Enlightenment.

VERNON,R., (1976), "The 'Great Society' and the 'Open Society': Liberalism in Hayek and Popper", Canadian Journal of Political Sciences, 9, 2 (1976).

VINER,J., (1927), "Adam Smith and Laissez Faire", Journal of Political Economy, 35 (April 1927), pp.198-232.

WAGNER, H.R.(ed.), (1975), On Phenomonology and Social Relations. Chicago: University of Chicago Press.

WALSH, G. AND LEHNERT, F. (eds.), (1967), The Phenomenology of the Social World. Evanston: Northwestern University Press.

WEIBULL,J.W., (1995), Evolutionary Game Theory, Cambridge, Mass & London: The MIT Press.

WEIBULL,J.W., AND J. BJÖRNERSTEDT, (1996), "Nash Equilibrium and Evolution by Imitation", IN: K. ARROW ET AL, The Rational Foundations of Economic Behaviour, London: MacMillan Press Ltd.

WEINTRAUB,E.R.(ed.), (1992), Toward a History of Game Theory, Durham and London: Duke University Press.

WILLIAMSON,O., (1975), Markets and Hierarchies: analysis and antitrust implications - a study in the economics of internal organization, New York: Free Press.

WINCH,P., (1958), The Idea of a Social Science - and its relation to Philosophy, London and Henley: Routledge & Kegan Paul, New York: Humanities Press.